The Dissertation Committee for David Russell Bell Certifies that this is the approved version of the following dissertation:

Modeling RNA, Protein, and Synthetic Molecules Using Coarse-Grained and All-Atom Representations

Committee:

Pengyu Ren, Supervisor

Ron Elber

Jeanne Stachowiak

Marcelo Behar

Modeling RNA, Protein, and Synthetic Molecules Using Coarse-Grained and All-Atom Representations

by

David Russell Bell, M.Sc.Eng, B.Sc

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2016

Dedication

To my family

Acknowledgements

I would like to thank my advisor Dr. Pengyu Ren for his support and pointing me in the right direction numerous times. We studied some fascinating problems and I am appreciative of the wisdom he shared.

I am grateful to the members of the Ren lab with whom I have collaborated and shared laughs with. This includes Chenfeng He, Sara Cheng, Jayvee Abella, Rui Qi, Dr. Zhen Xia, Dr. Qiantao Wang, Dr. Xiaojia Mu, Dr. Changsheng Chang, Zhifeng Jing, Matthew Harger, Dr. Chengwen Liu, Dr. Yue Shi, Rynne Ambrose, and Danny Dykstra.

I am fortunate and grateful to have worked under a grant from The Welch Foundation: grant F-1691 awarded to Dr. Pengyu Ren.

From IBM, I thank Ruhong Zhou for giving me the opportunity to work with his group for two summers. I greatly appreciate Seung-gu Kang for his excellent mentorship. I would also like to thank Seung-gu Kang and Jeff Weber for their collaboration with the quantum dot-SH3 and RNA aptamer projects. I would like to thank the rest of the group at IBM for the interesting discussions we shared.

Finally, I would like to thank my family for their support and putting up with me during my graduate study.

Modeling RNA, Protein, and Synthetic Molecules Using Coarse-Grained and All-Atom Representations

David Russell Bell, PhD

The University of Texas at Austin, 2016

Supervisor: Pengyu Ren

The aim of computational chemistry is to depict and understand the dynamics and interactions of molecular systems. In addition to increased comprehension in the physical and life sciences, this insight yields important applications to therapeutic design and materials science. In computational chemistry, molecules can be modeled in a number of representations depending on the molecular system and phenomena of interest. In this work, both simplified, coarse-grained representations and all-atom representations are used to model the interactions of RNA, cucurbituril host-guest chemistry, and cadmium selenide quantum dot binding to the Src homology 3 domain.

For RNA, a coarse-grained model was developed termed RACER (RnA CoarsE-gRained) to accurately predict RNA structure and folding free energy. After optimization to statistical potentials, RACER accurately predicted the structures of 14 RNAs with an average 4.15Å root mean square deviation (RMSD) to the experimental structure. Further, RACER captured the sequence-specific variation in folding free energy for a set of 6 RNA hairpins and 5 RNA duplexes, with a R^2 correlation of 0.96 to experiment.

The binding free energies of a cucurbituril host with 14 guests were computed using a polarizable force field and the free energy techniques of Bennett acceptance ratio and the orthogonal space random walk. The polarizable force field captured binding

accurately, yet unexpectedly, the orthogonal space random walk method converged slowly, albeit at still reduced computational expense to the Bennett acceptance ratio.

Lastly, the nanotoxicity effects of trioctylphosphine oxide coated cadmium selenide quantum dots are investigated with the model Src homology 3 protein domain in complex with its native proline rich motif ligand. With increasing quantum dot concentration, there is an increasing preference for the quantum dots to bind to the proline rich motif active site, inhibiting Src homology 3 function.

Table of Contents

List of Tables

List of Figures

Chapter 1: Dissertation outline and background of RNA modeling

1.1 Outline of dissertation

In this first chapter, I present the RNA folding problem and the need for high accuracy RNA structure prediction, the main focus of my work. I discuss the current methods used to predict RNA structure, including secondary and tertiary structure prediction. I then end this chapter by briefly introducing the RACER RNA model.

In chapter 2, I present recent progress made on the RACER RNA model. This includes the optimization of the model to capture folding free energy landscapes in addition to accurate structure prediction. I present RACER's capability to predict experimentally determined folding free energies with a correlation of R^2=0.98 while maintaining a structure prediction accuracy of 4.15 Å RMSD for a set of 14 experimentally determined structures.

In chapter 3 I present the computation of binding free energies for a set of 14 small-molecule guests binding to a cucurbituril host. The binding free energies were computed using two methods: (1) the Bennett Acceptance Ratio (BAR) method, and (2) the Orthogonal Space Random Walk (OSRW) method. The OSRW method is an enhanced sampling free energy computation method, which we found to compute results comparable to the standard BAR method but at a reduced computational expense. For the 14 small-molecule guests, we also compute several additional analyses, such as the conformational entropy of the guests rotating inside the host.

Lastly, in chapter 4, I present the interaction between the SH3 protein domain, PRM native ligand, and trioctylphosphine oxide (TOPO) coated CdSe quantum dots (QD). The QDs quickly aggregate in solution; hence, we studied the concentration dependence of QDs interacting with the SH3 domain. We found that with increasing

concentration, there was an increased preference to interact with the PRM binding site on the SH3 domain. The hydrophobic TOPO chains are largely responsible for this preference, as with increasing concentration, the QD CdSe cores are sequestered, while the TOPO chains become more surface exposed. Our work agrees with experiment that QDs exhibit dose-dependence toxicity, but this toxicity is heavily mediated by their surface coating.

1.2 RNA STRUCTURE MODELING

1.2.1 Problem

The central dogma of biology considers RNA largely as a passive molecule: RNA polymerase transcribes messenger RNA which is then translated into protein by the ribosome and transfer RNA. In terms of protein coding, this procedure is correct; however, RNA has extensive function outside of translation. A large part of this function pertains to genome regulation, where self-splicing and conformational changes dictate which genes are translated. The function of RNA is dependent on RNA structure, and the process of how RNA forms its structure is termed the RNA folding problem.

The RNA folding problem is rife with challenges. As soon as RNA nucleotides are transcribed from the polymerase, they begin to locally interact with previously transcribed nucleotides. As the synthesized RNA chain lengthens, long-range interactions occur in addition to local contacts. Upon RNA transcription completion, the RNA chain is able to fully sample conformations, breaking local interactions for favorable long-range interactions. Although RNA is able to fold in the absence of the polymerase, *in vivo*, RNA folds co-transcriptionally, sampling local and then long-range conformations as the RNA is synthesized[1]. For many RNAs, there is no unique folding procedure; rather multiple heterogeneous folding pathways (conformations sampled before native

structure) occur. Compounding the multiple folding pathways is that some of the alternative conformations are metastable and may persist for several hours. The timescale of RNA folding is likewise heterogeneous: for simple hairpin helices of ~10nt, 10-100μs is assumed reasonable, while the 195nt *Azoarcus* ribozyme folds in <50ms thanks to beneficial tertiary interactions[2,3], and perhaps the far end of the timescale is the 400+nt *Tetrahymena* ribozyme which can take hours thanks to long-lived misfolded intermediates[4,5]. This occurrence of metastable non-native RNA structures has been dubbed the 'alternative conformer hell'[6,7] as it is exceedingly frustrating for structural biologists using techniques such as x-ray crystallography, NMR, and cryo-em.

Given that experimental approaches to determining RNA structure are so challenging, there has been substantial development of theoretical models to predict RNA structure. Similar to Levinthal's paradox[8] for protein folding, RNA folding is utterly too expensive to exhaustively search every conformation. RNA structure prediction models must hence simplify the folding procedure to still capture accurate RNA structure in a practical amount of time. In response to the need for RNA structure as well as the prediction challenges, the Ren lab has developed the RACER model to predict RNA structure based on physical interactions. The physics of RNA folding will be discussed next followed by a short review of other structure prediction programs, and then the RACER model will be presented.

1.2.2 Physics

The main interactions responsible for RNA structure are those of electrostatics, hydrogen bonding, pi-pi stacking, and solvation. The phosphate backbone of RNA holds a charge of -1*e* per nucleotide. This leads to strong backbone-backbone repulsion between phosphate groups. Several species of ions particularly Mg^{2++} in solution and

bound to the RNA act to screen this repulsion. Mg^{2++} ions may bind to RNA in several different modes[9]. Hydrogen bonding in RNA occurs through base pairing interactions, a result of the polarity of the nucleobase heavy atoms O, N, and C. If one considers the RNA base to be a flat aromatic molecule with four edges, one side is occupied by the RNA backbone; the other three sides (Watson-Crick, Hoogsteen, and sugar edges) are capable of forming hydrogen bonds. The canonical base-pairs, GC and AU, form 3 and 2 hydrogen bonds respectively. The extra hydrogen bond on the GC basepair is observed to be exceedingly stable. Non-canonical basepairs such as AA also occur in physiological RNAs. The hydrogen bonding between GG in G quadruplexes results in high stability structures. Pi-Pi stacking is a result of the delocalized π electrons on the aromatic bases, which contribute to attraction between nucleobase planes. This attraction is termed base stacking and is observed to be as stabilizing as base pairing. Besides adjacent nucleobases participating in stacking interactions, nucleobases may base-stack with nucleobases on the other strand of a base-paired region, referred to as cross-stacking. Solvation effects for RNA folding are not as prominent for RNA folding as they are in protein folding. However, the charged backbones have a strong preference to be solvent exposed, while the nucleobases hold less preference to interact with solvent. Again, this effect is minor in most instances.

1.2.3 Secondary Structure Prediction

Beyond sequence (primary structure), a highly simplistic representation with surprisingly powerful results is that of secondary structure representation. In secondary structure representation, RNA base-pairing is mapped out, so that helices and loops are evident. From observing ribosomal RNA secondary structure, Woese, Gutell, Cannone

and coworkers have drawn some astounding conclusions, including the separation of the archaebacteria and eubacteria domain.

The most popular secondary structure prediction program, Mfold, uses dynamic programming to predict RNA secondary structure. Mfold works by comparing two different base-pairing conformations and taking the lowest free energy conformation to next compare with another base-pairing combination. Mfold maintains limitations, such as only ≈20-60% accuracy on the 16S rRNA[10], as well as over-stabilizing helices. However, if considered appropriately, Mfold may serve as a nice approximation for RNA secondary structure, especially given the expense of determining secondary structure experimentally. Free energies are assigned to neighboring base-pairs, with these termed nearest-neighbor free energies stemming from a compendium of melting free energies collected by Turner and coworkers[11]. Melting free energies are determined by measuring the absorption profile as a sample of double-stranded RNA is heated. Single- and double-stranded RNA have different absorption spectrums, resulting in a sigmoidal curve as the RNA molecules transition from being mostly double stranded to mostly single stranded. It is important to note that the high temperature after melting prevents the single stranded RNA from forming hairpins and long-lasting tertiary interactions.

A large variety of other RNA secondary structure programs exist and are widely used. Some programs follow the dynamic programming scheme of Mfold. Others are capable of predicting pseudoknot structures. Although secondary structure prediction programs are powerful and heavily used by researchers, they remain limited to 2-dimensional structure representations. 2-dimensional representations are insufficient for understanding which regions are solvent exposed or capturing specific binding modes of ligands/ions. For reasons such as these, 3-dimensional RNA structure is needed for future understanding and application of RNA biology.

1.2.4 Tertiary Structure Prediction

3-dimensional RNA structure prediction methods typically use either RNA fragments to assemble 3-D structures, topological methods, or Monte carlo/Molecular Dynamics 'physics-based' sampling to build 3-D structures. In fragment assembly methods, the RNA sequence or secondary structure is divided into small fragments, such as the 8nt or shorter Nucleotide Cyclic Motifs from MC-Sym/MC-fold[12]. The fragments are collected from solved crystal structures and then placed into a database. New RNA structures are then built from the fragments, with a scoring function determining the optimal structure. In many fragment assembly programs, RNA secondary structure is first predicted which narrows down the selection of possible fragments. Also, some programs allow for relaxation/minimization of the fragments once the RNA is built so as to allow predictive flexibility from the constituent fragments. Fragment assembly methods accuracy is proven by their performance in *RNA Puzzles*[13,14], a competition amongst the RNA structure prediction community to blindly predict RNAs from sequence, with the experimental RNA structure being released after the entries are submitted and then the programs are scored.

Topological methods to predict RNA structure have been spearheaded by Schlick[15-20] as well as a few others[21-24]. In these methods, RNA is depicted as nodes and vertices with applications ranging from RNA compaction to prediction of novel motifs.

Physics-based RNA structure models aim to predict RNA structure from the physical interactions that the molecule encounters. Most physics-based models depict the RNA at a coarse-grained level, with a few parameterized pseudoparticles representing the nucleotides rather than the all atom structure. These pseudoparticles are typically modeled using a standard molecular mechanics force field, so that bonded pseudoparticles experience stretching, bending, and torsion forces. Nonbonded

interactions vary depending on the level of pseudoparticle representation. Effective non-bonded interactions typically follow a vdW-like functional form where at the long-range (~5-20 Angstroms) the pseudoparticles experience attraction, while at the short range, the particles experience repulsion. Other interaction terms that tertiary structure models use are electrostatics, as well as separate base pair and base stacking interactions. Ultimately, non-bonded interaction composition depends on the level of coarse-grained pseudoparticles depicted by the model.

1.2.5 RACER Structure Prediction Model

The RACER structure prediction model is a coarse-grained physics based RNA structure prediction model developed by the Ren lab. Previously, the model has been shown to predict RNA crystal structures accurately and efficiently[25,26]. The RACER model consists of 5-pseudoparticles per nucleotide, with one pseudoparticle on the phosphate group, one pseudoparticle on the C4' of the ribose group, and three pseudoparticles on the nucleobases capturing nucleobase planarity (see Figure 1.1). RACER models water and salt implicitly, so it is only the RNA pseudoparticles that are simulated. Using RACER, we are able to run 1μs of molecular dynamics simulations in about one day on one cpu core for a 10nt hairpin. This simulation time is substantially fast and is a notable strength of the RACER model.

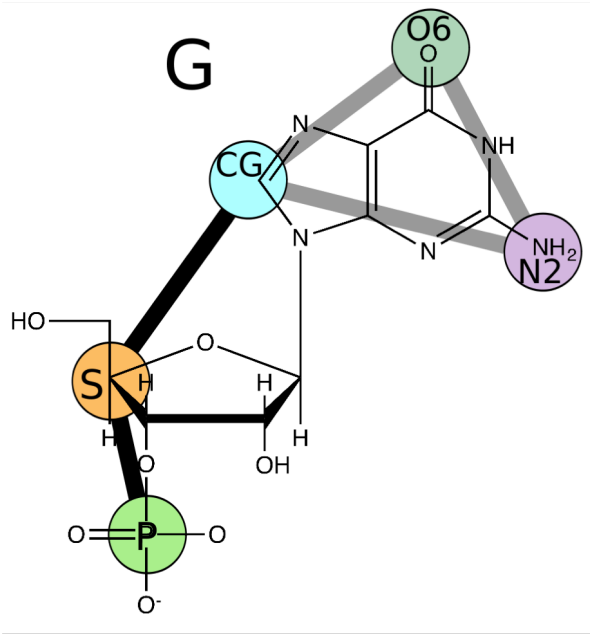


Figure 1.1: RACER model structure overlaid onto the all-atom structure of a guanine nucleotide.

In the next chapter, I discuss the most recent work on the RACER RNA model. This work involved re-optimizing the model to updated statistical potentials and then fitting the model to melting free energy data and pulling experiment data in addition to structure prediction. The correlation to experimental free energies is rather strong, with $R^2 = 0.98$. RACER structure prediction capability became slightly worse, with an average RMSD of 4.15Å to experiment, compared to the previous RMSD value of 3.31Å. Further discussion of model capability and what we modified in the model is presented in the next chapter. Ultimately, now the RACER model holds application towards RNA free energy landscapes, which are highly sought after in the RNA folding field, with the expense of reduced structure prediction capability.

Chapter 2: Capturing RNA folding free energy with coarse-grained molecular dynamics simulations[1]

2.1 INTRODUCTION

RNA serves important and diverse functions inside the cell. In 1981, Thomas Cech and colleagues observed self-splicing RNA in a 26S rRNA precursor[27,28]. In 1983, Sidney Altman found that ribonuclease P could cleave tRNA in the absence of protein[29]. In 2002, it was discovered that even mRNAs could bind small metabolites and regulate protein expression[30-32]. Today, RNA is recognized as extensively active, with roles in regulating genes, preparatory cleavage, metabolite sensing, and immune response. RNAs achieve this diverse activity through intricately regulated structure, with catalytic RNAs such as riboswitches maintaining highly conserved functional regions[33,34].

RNA chemistry and the need for accurate structures. RNA structure is a challenge to determine experimentally because it can fold into many different structures. For example, during RNA transcription, synthesized RNA regions fold locally[1], sampling hairpins and short-range motifs. After transcription completes, RNA molecules are able to fold completely and sample long-range interactions[35]. With the numerous structures available for RNA to fold into, long-lived misfolded RNA intermediates often occur[5,36,37]. In addition, heterogeneous folding pathways exist for the same RNA sequence[38-44]. As a result, RNA has a highly dynamic folding landscape, which is challenging to capture using techniques such as x-ray crystallography and NMR spectroscopy[6,7]. Further, due to only recent interest in the diversity of RNA function in biology, there is a deficiency in available RNA experimental structures. However, RNA structure is key to understanding its function and for development of RNA-based applications. Due to the lack of available

[1]Large portions of this chapter are based on the work: Bell, DR., et al. Capturing RNA Folding Free Energy with Coarse-Grained Molecular Dynamics Simulations. *Sci. Reports*. 2016

experimental structures of RNA, computational models of RNA are vital to predict RNA structures.

Secondary structure methods for RNA. Currently, there are a variety of structure prediction methods available to elucidate RNA structure. Secondary structure prediction methods predict base pairing contacts for a given RNA sequence[45]. If homologous sequences exist, comparative sequence analysis[46-49] remains the most accurate secondary structure technique. One of the most popular secondary structure prediction methods is dynamic programming. Using nearest neighbor energies[11] and the sequence of the RNA, dynamic programming methods, such as Mfold[50,51] or ViennaRNA[52-54], exhaustively compare and build secondary structures to achieve the minimum free energy structure.

However, dynamic programming schemes face certain limitations[10], such as difficulty predicting pseudoknot structures. Various secondary structure programs[55-57] have been developed to predict the folding of these structures. Recently, it has been shown that incorporating results from the experimental method SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension)[58] can moderately increase accuracy of secondary structure prediction[59-64]. Despite its utility, secondary structure prediction is ultimately limited to 2-D base paired RNA structures. For RNA based therapeutics and *de novo* design, 3-D RNA structure must be determined.

3-D structure prediction models. Tertiary or 3-D structure prediction methods use template, graph theory, and physics based modeling to sample and predict relevant 3-D RNA structures[65,66]. Template based modeling uses predefined, small motifs to assemble RNA structures from their sequence. Template based models include the MC-Fold/MC-Sym pipeline[12], BARNACLE[67], RSIM[68], 3dRNA[69], RNAComposer[70], Vfold[71-74], RNA-MoIP[75] and FARNA/FARFAR[76-78] available in the Rosetta package[79]. Similar to

template based modeling, ASSEMBLE[80] and RNA2D3D[81] use homologous RNA structures to predict the new RNA structure (with manual refinement available). In graph theory techniques, RNA is depicted topologically to build RNA structure, this improves sampling and even allows for creation of novel RNA motifs. Graph theory techniques[18] are utilized by RAG/RAGTOP[16,17,19,20] and others[21-24]. In physics based methods, the RNA is built from sequence into a 3D structure, and these 3D RNA structures are sampled using Monte Carlo or Molecular Dynamics (MD) protocols. Due to the high charge density of RNA and the associated large computational cost to sample structures, many tertiary structure models use coarse-grained representations of RNA[82].

In coarse-grained (CG) models, atomic sites are grouped together and represented as a "bead" or pseudoatom. Typical coarse-grained models depict a few pseudoatoms per nucleotide. This results in a reduction in the degrees of freedom and lowers the simulation cost of the model, as compared with simulating the all-atom structure. Physics based coarse-grained models with one pseudoatom per nucleotide include YAMMP/YUP[83,84], an adaptable user input required model, and NAST[85,86], which assumes ideal helices from secondary structure and uses MD and clustering to build loops. iFoldRNA[87,88], Denesyuk et al.[89,90], and TOPRNA[91-93] use three pseudoatoms per nucleotide to depict phosphate, sugar, and nucleobase groups. iFoldRNA uses discrete Molecular Dynamics and replica exchange Molecular Dynamics to sample structures, with non-bonded parameters decomposed from nearest neighbor energies. Similarly, the model by Denesyuk et al.[89,90] derives its parameters from nearest neighbor energies and experimentally determined structures. TOPRNA captures effects of secondary structure constraints on loop conformations and free energies. HiRE-RNA[94-96] depicts six-seven pseudoatoms per nucleotide with five pseudoatoms along the backbone. SimRNA[97,98], Bernauer et al.[99], as well as the previous generation and current RACER model

studied[25,26,100] , all represent RNA with five pseudoatoms per nucleotide. SimRNA uses a Monte Carlo sampling algorithm with parameters from statistical potentials. The model by Bernauer et al. similarly uses statistics from high-resolution crystal structures for parameterization yet also derives all-atom potentials for structure refinement.

The RACER RNA Model. The CG RNA model RACER (RnA CoarsE-gRained) developed and applied in this work is a physics-based model, derived from RNA structural statistics, refined using RNA thermodynamics, and applied in molecular dynamics simulations of folding and complexation of RNAs. In the results section, we first introduce the potential energy functions used in the RACER model, with a focus on the newly implemented effective vdW potential. Second, we demonstrate how RACER parameters were optimized using statistical potentials derived from PDB statistics. Additionally, we provide motivation for modeling RNA as a 1D molecule and the associated 1-D correction we made to the non-bonded PMFs. Third, we show how we validated RACER using simulated annealing simulations for RACER structure prediction capability and generation of funnel free energy landscapes. Fourth, we apply RACER to generate folding free energy predictions for a testing set of RNA hairpins and duplexes, and we compare our results to experiments. In the discussion section, we summarize the changes made to the RACER model and emphasize RACER's ability to capture folding free energies and to predict structures. In the methods section, we show (1) the ability of RACER to map between all-atom and coarse-grained representations for use in multiscale simulations, (2) details on the folding free energy calculations, and (3) implementation instructions for those wishing to use RACER.

2.2 RESULTS

2.2.1 Model

Potential energy functions. The total potential energy function of the RACER model includes bond stretching, angle bending, torsion, effective vdW, hydrogen bonding, and electrostatics, labeled as E_{bond}, E_{angle}, $E_{torsion}$, E_{vdW_eff}, E_{hb}, and E_{ele} respectively (see Eq. 1). The RACER model is currently implemented in TINKER[101]. In RACER RNA nucleotides consists of 5 pseudoatoms per nucleotide, with a total of 9 pseudoatom types (shown in Figure 2.1). The RACER model used here differs from previous publications[25,100] in that we employ a novel effective vdW potential to better capture the short-range non-bonded interactions among the pseudoatoms, which we found to be essential for correctly capturing the folded state. As a result, we had to re-parameterize the other non-bonded contributors including the electrostatics and hydrogen bonding potential.

$$E = E_{bond} + E_{angle} + E_{torsion} + E_{vdW\_eff} + E_{hb} + E_{ele} \quad (1)$$

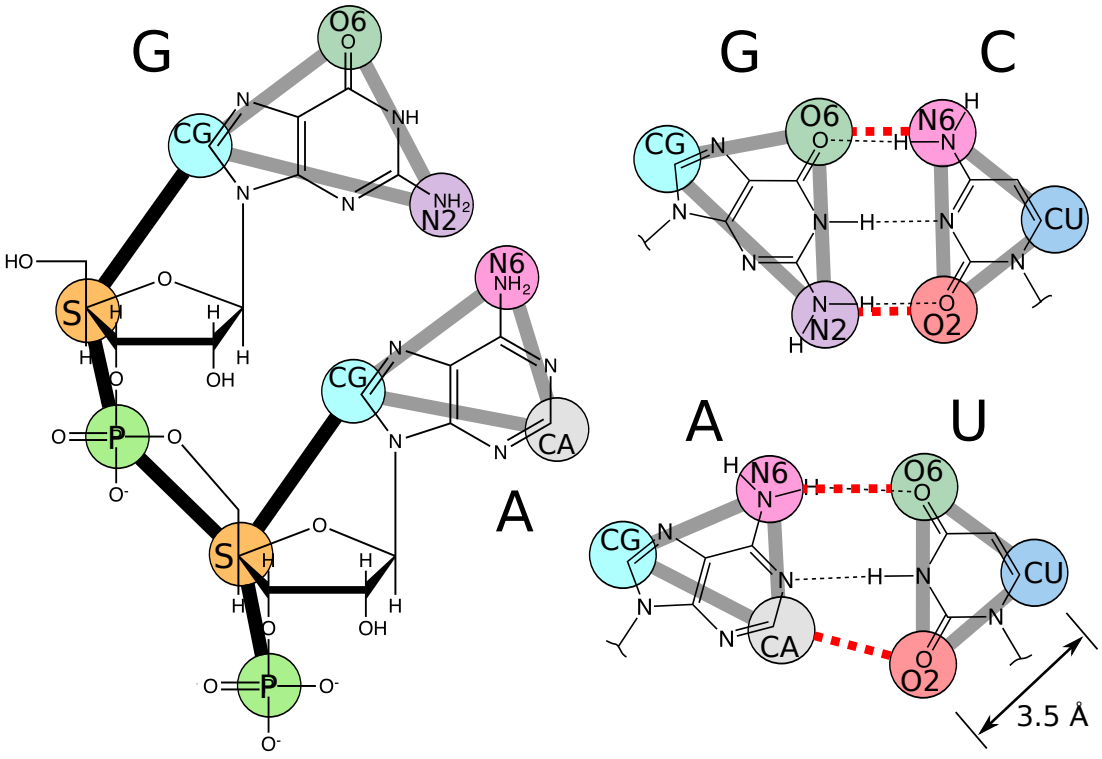


Figure 2.1: RACER model pseudoatoms overlapping all-atom structure. RACER bonds are shown in bold lines. (Left) nucleobases with backbone. (Right) GC and AU basepairs with hydrogen bonds shown in red dashed lines. Scale bar is shown in lower right.

Bonded Potential Energies. The potential energy functions which retain the same functional form between the previous model and RACER are the bonded potential energy functions. Bond and angle potentials are represented by harmonic terms: $E_{bond} = k_{bond}(b - b_o)^2$ and $E_{angle} = k_{angle}(\theta - \theta_o)^2$. The torsion potential of Eq. 2 uses the first 3 terms of a Fourier series expansion for the torsion potential, where ϕ is the torsion angle, and k_n and δ_n are the spring constant and phase angle of expansion term n.

$$E_{torsion}(\phi) = \sum_{n=1}^{3} k_n(\,1 + \cos(n\phi - \delta_n)) \tag{2}$$

Improved Effective vdW Potential. The RACER model includes a newly implemented effective potential (vdW_{eff}) that significantly improves the fit of RACER to non-bonded statistical potentials. In the previous model[26] the vdW-like non-bonded potential was modeled using a Buckingham function. However, this was found to significantly overestimate repulsion at short distances when compared with statistical potentials. The new effective vdW_{eff} potential (Eq 5) allows for tuning the repulsion at short distances through a third parameter γ, enabling a closer fit to the statistical non-bonded potential of mean force (PMF) (Figure 2.S1).

The vdW_{eff} does not represent the true vdW interaction, but rather the potential of mean force between a pair of pseudoatoms. However, based on statistical potentials, the non-bonded interactions between most pairs of pseudoatoms we sampled exhibited vdW potential-like behavior. The new functional form for vdW_{eff} potential taken from ref [102] is shown in Eq. 5, where ε is the minimum well depth and σ is the distance of minimum energy, and γ is a parameter allowing for fine-tuning of the slope of the short-range interaction. Figure 2.S1a presents a comparison between the vdW_{eff}, Lennard Jones, and Buckingham potentials while Figure 2.S1b-d show the effects of the three parameters σ, ε, and γ on the vdW_{eff} potential. The combining rules for unlike pseudoatom types i and j in the vdW_{eff} potential are: $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$, and $\gamma_{ij} = (\gamma_i + \gamma_j)/2$.

$$E_{vdW\_eff} = \frac{2\varepsilon}{1 - \frac{3}{\gamma+3}} \left(\frac{\sigma^6}{\sigma^6 + r^6} \right) \left[\frac{3}{\gamma + 3} e^{\gamma\left(1 - \frac{r}{\sigma}\right)} - 1 \right] \quad (5)$$

Hydrogen Bond and Electrostatics Energies. The hydrogen bond (Eq. 3) and Debye-Huckel electrostatics (Eq. 4) potential energy terms are of the same form as used previously. However, we reparametrized the hydrogen bond and Debye-Huckel potentials with the introduction of the new vdW_{eff} term. In the hydrogen bond potential, $\varepsilon_{hb,max}$ is the maximum potential found at the hydrogen bond equilibrium distance $\sigma_{hb,eq}$. $\left|\overrightarrow{r_{ji}}\right|$ is the magnitude of the vector from atom j to atom i, while $\alpha_k = 2\left(\theta_i + \theta_j\right) - \pi$ is a directional component with θ_i and θ_j defined in Figure 2.S2. For hydrogen bond parameterization, the maximum potential $\varepsilon_{hb,max}$, was increased from 0.5 kcal/mol to 2.0 kcal/mol. Other hydrogen bond parameters including equilibrium distance $\sigma_{hb,eq}$ of 2.9 Å and cutoff of 6 Å (base edge) remain the same as the previous model. Hydrogen bond potential energy is only computed for GC and AU pairs. For Debye-Huckel Eq. 4, q_i is the charge of atom i, r_{ij} is the distance between atom i and atom j, D is the dielectric constant, and ξ is the Debye length. A dielectric constant D of 25 was determined to be optimal under the new model potential, compared to 78 from the previous model. In depth discussion of Debye-Huckel and hydrogen bond optimization can be found in the SI.

$$E_{hb} = -\frac{\varepsilon_{hb,max}}{2}\left(1 - \cos(\alpha_k)\right)\left(\frac{\sigma_{hb,eq}}{\left|\overrightarrow{r_{ji}}\right|}\right)^3 \tag{3}$$

$$E_{ele} = \frac{q_i q_j}{4\pi D} {r_{ij}}^{-1} e^{-r_{ij}/\xi} \tag{4}$$

2.2.2 Model Improvement and Parameterization

Statistical potentials. The premise of our parameter optimization was to fit to both RNA structure and experimental free energies. First, we updated model statistical potentials from experimentally determined crystal structures. We downloaded all available Protein Data Bank (PDB, http://www.rcsb.org/) RNA structures as of 02/10/15

(excluding RNA-protein and RNA-DNA combination structures) totaling 1100 entries Our previous model fit to statistical potentials used approximately 668 structures. For RACER, our updated parameterization includes an additional ~400 structures, which led to various modifications in the potentials. The method of statistical potentials involves fitting energy functions to statistically derived potential of mean force (PMF) curves. The PMFs are determined by taking the probability distribution $P(r)$ of occurrences from the PDB structure set and then extracting the free energy $G(r)$, $G(r) = -k_B T \ln\left(\frac{P(r)}{ref}\right)$ with the reference distribution ref setting the minimum interaction at 0 kcal/mol. Harmonic terms for valence potential, including bond and angle potentials, only required minor adjustments from previous values to fit to the updated PMF curves (most within 2% to the previous model[26]).

1-D PMF for RNA. In this work, we determined that modeling RNA as one-dimensional rather than a three-dimensional, isotropic molecule is more appropriate when extracting the statistical potentials from PDB structures. This choice is justified as there is an abundance of short, linear helices found in PDB structures of RNA. Additionally, folded RNA typically forms prolate ellipsoids[103]. Similarly, in the PDB structure of 16S rRNA more than half of the nucleotides are base paired[46]. Therefore, treating RNA as a one-dimensional molecule for capture of local interactions is not unreasonable. Additionally, 3D PMFs are more appropriate for systems with isotropic distance distributions, such as molecular liquids [104-106] and proteins[107-109].

Our motivation for modeling RNA as a 1D molecule came from the observation of divergence of 3D radial distribution functions (RDF) at distances greater than 10 Å, and as a result the potential of mean force (PMF) that was derived from the RDF did not converge to 0 at large separation (see Figure 2.2). However, when 1D radial distribution functions were used the PMF asymptotically approached zero for long distances (see

Figure 2.2), reinforcing the discussion above. The main difference between 3D and 1D RDFs is the normalization factor. For 3D RDFs, normalization is done over volumetric shell $4\pi r^2 dr$, whereas 1D RDFs normalizes over an incremental distance, dr.

Specifically, the non-bonded PMF is evaluated via Boltzmann inversion as $G(r) = -k_B T \ln(g(r))$ where $g(r)$ is the radial distribution function, normalized probability function discussed above. When treating RNA as a 3D isotropic molecule, the 3D RDF, as was done previously,[25,100] is given by $g(r) = n_{ij}(r)/\left[\left(N_i N_j/V\right)4\pi r^2 dr\right]$, where $n_{ij}(r)$ is the number of atom type j at distance r from atom type i, N_i and N_j are the total number of i and j atoms respectively, and V is the volume of the system. Now we treat RNA as a "1D", linear molecule to more adequately parameterize the vdW_{eff} potential, and the RDF becomes $g(r) = n_{ij}(r)/\left[\left(N_i N_j/V\right)dr\right]$.

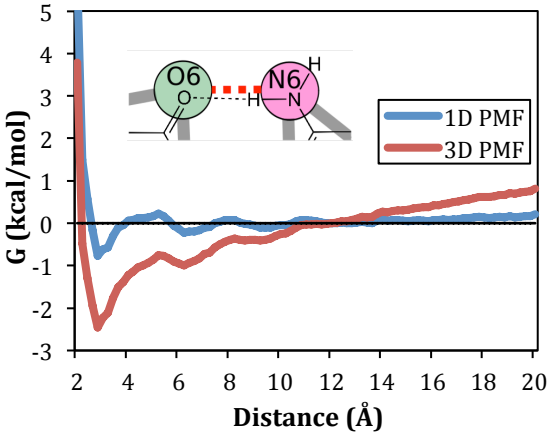


Figure 2.2: Comparison of a 1-D and 3-D statistical potential PMF, computed from a 1-D and 3-D radial distribution function respectively. Note how the 3-D PMF continues to diverge at long distances whereas the 1-D PMF falls off to zero. This statistical potential is that of RACER base pseudoatoms O6-N6.

2.2.3 Structure Prediction

Folding RNA by simulated annealing. We tested RACER with simulated annealing simulations to (1) validate that RACER can accurately fold experimentally

determined RNA structures and to (2) ensure the native structure has the lowest energy on its energy landscape. We ran simulated annealing simulations on a testing set of 14 RNAs, duplexes and hairpins, that have known experimentally determined structure and free energies[100]. RACER is able to predict, from simulated annealing, ten out of fourteen RNA molecules with RMSD < 5 Å, and six RNA molecules with RMSD < 2.5Å. The average RMSD between the predicted lowest-energy structures and native structures is 4.15Å. This average RMSD is slightly worse than our previously published average RMSD of 3.31Å, however, now our model has the capability to predict free energy landscapes of RNA in addition to structure prediction.

The simulated annealing protocol involved running MD sequentially for 5ns at temperatures in order of 298(K), 400, 1000, 900, 800, 700, 600, 500, 400, 298K, for a total simulation time of 50ns, with structures saved every 10ps. Given the high temperatures used, we used a 1fs time step for annealing simulations. Results for structure prediction using simulated annealing are given in Table 2.1. These predicted RMSD values are calculated between PDB structures and the minimum potential energy structures found by RACER.

Table 2.1: Predicted (minimum energy) RMSD values compared to Protein Data Bank (PDB) structures from simulated annealing.

PDB ID	157D	1AL5	1DQF	1F5G	1I9X	1KD5	1LNT	1QCU	1ZIH	2AO5	2JXQ	2K7E	353D	472D	**Avg.**
RMSD (Å)	1.45	1.31	3.50	7.75	4.54	8.05	7.67	2.00	4.88	1.84	1.13	8.04	2.47	3.44	4.15 ± 0.72

Energy landscapes. Analyzing the energy landscapes of the 14 RNAs in our training set was an important part of our optimization. RNAs are complex molecules that may adopt stable and long lived misfolded structures. However, it is assumed the final native structures, at least in vitro, should have the lowest free energy for the given

environment[110]. Here we generate for RNA a large number of unfolded structure by simulated annealing, followed by energy minimization of each the structure. The energy and RMSD (with respect to the native structure) are used to characterized the energy landscape. The energy-RMSD landscapes for all 14 RNAs are given in SI, Table 2.S1.

The energy vs RMSD landscapes for all 14 RNAs show clear "funnel" shapes skewed toward the native structure. As examples, we present the energy landscapes for two favorably predicted structures (157D and 1AL5) in Figure 2.3a-b, and the energy landscapes for two unfavorably predicted structures (1F5G and 1KD5), where the lowest energy structures have large RMSD in Figure 2.3c-d

Representative energy landscapes for PDB ID: 157D, 1AL5, 1F5G, and 1KD5 are shown in Figure 2.3. RACER predicted structures 157D and 1AL5 agree well with experiment (inset in Figure 2.3a-b), but the RACER predicted structures for 1F5G and 1KD5 have collapsed into torus-like structures, with very little backbone twist (inset in Figure 2.3c-d). A possible explanation for this observed behavior is the presence of non-canonical base pairing in 1F5G, which has a G-G basepair, and 1KD5, which has 4 basepair mismatches. At this time, RACER only captures AU and GC hydrogen bonds; however, noncanonical hydrogen bonds are a target for future development.

Additionally, energy landscapes allow us to identify possible meta-stable intermediates, which are high-RMSD (~10 Å) "local" funnels observed in plots for 1AL5 and 1QCU. The meta-stable structure of 1AL5 at the local minimum, shown in Figure 2.S3 resemble toroidal structures observed for 1F5G and 1KD5, but for 1QCU an extended, base stacking meta-stable structure is observed. In our model, the vdW_{eff} energy is mainly responsible for these meta-stable structures: it is ~-50 kcal/mol more favorable than in the predicted global-minimum structures, which are very close to the experimental structures.

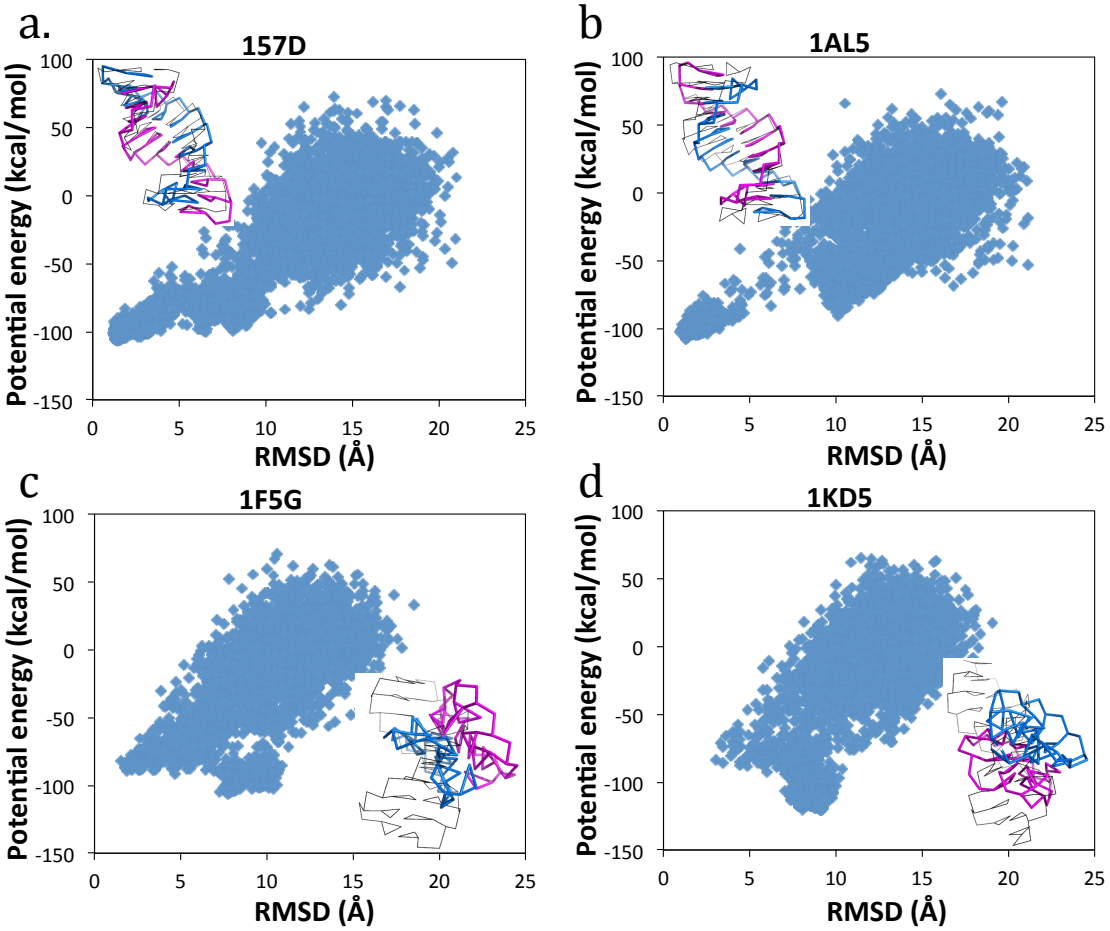


Figure 2.3: Representative energy landscapes from annealing for two RNAs that are accurately predicted: (a) 157D, (b)1AL5, and two RNAs that are poorly predicted: (c) 1F5G, and (d) 1KD5. For each RNA, the RACER minimum free energy structure is shown in blue and magenta sticks aligned to the PDB structure shown in black lines. Five thousand structures over 50ns are shown for each RNA; each structure is energy minimized before plotting. Note the funnel toward low energy and low RMSD structures. The RMSD of lowest energy structure for 157D is 1.45Å, 1AL5 is 1.31Å, 1F5G is 7.75Å, and 1KD5 is 8.05Å.

In the process of validating and optimizing our model by energy landscape analysis, we notice the importance of a dedicated hydrogen bond potential for base paring, as the vdW_{eff} potential is not well suited for distinguishing between base stacking

and base pairing interactions[111]. The hydrogen bond potential allows for directional base pairing and helps in separating the base stacking and base pairing interactions effectively.

2.2.4 Equilibrium Pulling Simulations

Experimental free energies. To test RACER, we focused on capturing experimental melting free energies of canonical helices[112] and hairpins[113]. We used RACER to perform equilibrium pulling simulations, and we compared free energy differences to two sets of experimental thermodynamic data: RNA melting free energies from Turner and coworkers[11] and folding free energies from single molecule force experiments. Five hairpins of size 10, 10, 12, 14, and 18nt and five duplexes of size 6, 8, and 10 base pairs were selected from melting free energy experiments, and the TAR RNA hairpin was chosen to compare RACER to single molecule force experiments. Hairpin sequences 30, 11, 19, 33, and 47 from the supplementary information of [113] referred to here as h1, h2, h3, h4, and h5 and duplex sequences 35, 48, 71, 78, and 90 of [112] referred to here as d35, d48, d71, d78, and d90. TAR is a 52nt, 21 bp hairpin with two internal loops.

In melting free energy experiments, a solution of RNAs of known sequence are heated while measuring UV absorption. As helical and single stranded RNAs absorb light at different wavelengths, the absorption will change over heating as the RNA denatures. By fitting a curve to absorption vs temperature the melting free energy can be determined.[114-116] Turner and co-workers have published a compendium of melting free energies for small RNA motifs and structures using nearest neighbor energy parameters and RNA secondary structure prediction[11,112,117]. Additionally, we compared our model to RNA single molecule force experiments.

In single molecule force experiments, folded RNA molecules are unfolded by mechanical force using techniques such as optical tweezers or atomic force microscopy. Using the end-to-end extension as a reaction coordinate, the free energy of unfolding can be determined from position vs. time data. A recent single molecule research study of the trans activation response (TAR) element of HIV extracted the free energy of folding at zero force under the assumption of the worm-like chain model[118]. Here we study the same TAR RNA as used in the single molecule force experiments.

Melting and pulling experiments for all RNAs were simulated by umbrella sampling simulations pulling the RNAs apart from their ends (see Figure 2.4 for example simulation setup showing end-to-end reaction coordinate). The folding free energy values were then computed using the Weighted Histogram Analysis Method (WHAM) software distributed by Alan Grossfield[119]. Details of these simulations and computations are included in the Methods section. Although exact energy landscapes at equilibrium for both TAR and melting free energy helices are unknown, the free energy difference between unfolded and folded states (ΔG) can be determined using Eq. 5[120] and then compared to results from experiments. In Eq. 5, V_{ref} is the standard state reference volume (1660 A^3/molecule), $\Delta\omega$ is the free energy difference from the PMF (ω_{folded} - $\omega_{unfolded}$), r is the distance coordinate between RNA ends, and Ω is the orientation coordinates for the RNA. In this work, $d\Omega=4\pi r^2$ and $dr=0.25$, the bin size of the WHAM calculation.

$$\Delta G = kTln\left(\frac{4}{3}\pi^2 V_{ref}\right) - kTln\left(\int drd\Omega e^{-\frac{\Delta\omega(r,\Omega)}{kT}}\right) \qquad 5$$

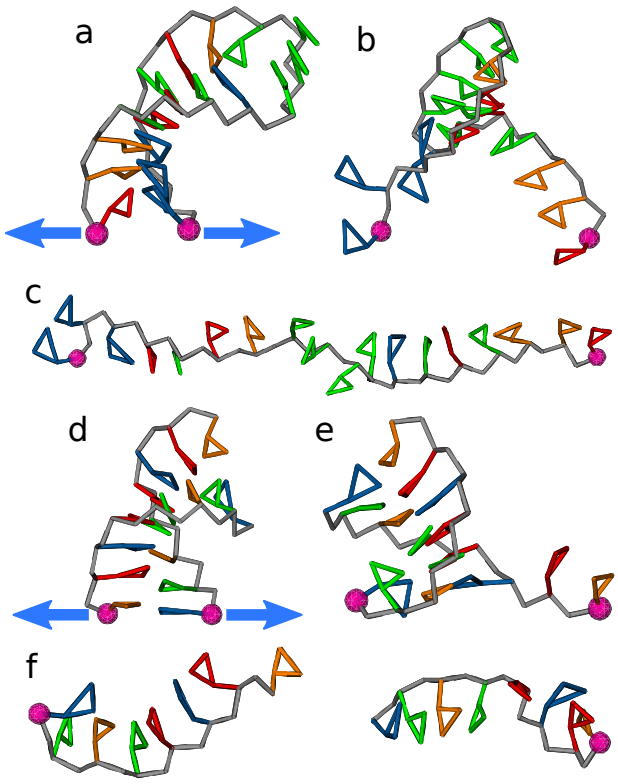


Figure 2.4: Pulling simulation setup of Hairpin h3 (a-c) and duplex d78 (d-f). The RNAs were pulled along the reaction coordinate of end-to-end extension (marked by large magenta spheres) using umbrella simulations. The magenta spheres at the strand ends represent the sugar pseudoatoms that were restrained in umbrella simulations. (a, d) native end-to-end extension (b, e) partially denaturing extensions and (c, f) and unfolded/melted extensions. Note that in the folded structures, base stacking and base pairing interactions exist, while in unfolded or melted structures, only base stacking interaction exists. Gray bonds are backbone atoms while red, orange, green, and blue bonds are A, C, U, and G nucleobases respectively.

Unfolding free energy from RACER MD simulations. The free energies computed from equilibrium pulling MD simulations (WHAM) using RACER are in excellent agreement with experimental measurements, with a correlation coefficient (R^2) of 0.98 for 7 RNAs tested (Table 2.2 and Figure 2.5). For additional comparison, we also

included the melting free energies from Mfold, a widely-used secondary structure prediction program that has been parameterized using the experimental melting thermodynamic data (Mfold predicted structures are shown in Figure 2.S4). The unfolding free energies evaluated by RACER and Mfold[51] are presented in Table 2.2 along experimental values and the length of each MD simulation. The correlation plots for RACER and Mfold show both models have R^2 correlation coefficients of 0.98. However, Mfold over predicts the stability of the TAR hairpin and the duplexes. Note that RACER is a 3D particle based physical model developed for molecular dynamics simulations, whereas Mfold is a 2D structure prediction program. In RACER we explicitly compute the entropy contributions to the free energy through molecular dynamics sampling.

Table 2.2: Unfolding free energy values for RNAs from experiment (Expt.), Mfold predicted, and RACER predicted. Molecule length in nucleotides or basepairs and the simulation length per window are also shown. Error is take from a Monte Carlo bootstrap error analysis as implemented in the WHAM program by Grossfield[119].

Hairpin	Length (nt)	Expt. ΔG (kcal/mol)	Mfold ΔG (kcal/mol)	RACER ΔG (kcal/mol)	Length per window
h1	10	-3.5 ± 0.3[113,121]	-5.3	-4.2 ± 0.06	1 μs
h2	10	-0.3 ± 0.1[113,122]	+0.9	-2.6 ± 0.19	1 μs
h3	18	-8.2 ± 0.2[113,123]	-8.4	-6.9 ± 0.28	1 μs
h4	12	-4.4 ± 0.2[113,124]	-3.4	-5.9 ± 0.22	1 μs
h5	14	-2.2 ± 0.08[113,125]	-2.2	-5.7 ± 0.24	1 μs
TAR	52	-21.5 ± 4.3[118]	-31.3	-19.1 ± 1.39	0.1 μs
Duplex	Length (bp)				
d35	6	-7.56 ± 0.3[112]	-11.4	-7.2 ± 0.22	1 μs
d48	6	-4.95 ± 0.2[112]	-9.8	-6.5 ± 0.22	1 μs
d71	8	-12.32 ± 1.2[112]	-17	-12.3 ± 0.25	1 μs
d78	8	-10.11 ± 0.4[112,114]	-14.7	-9.1 ± 0.27	1 μs
d90	10	-12.69 ± 0.5[112]	-18.1	-11.9 ± 0.30	1 μs
					Total: 534 μs

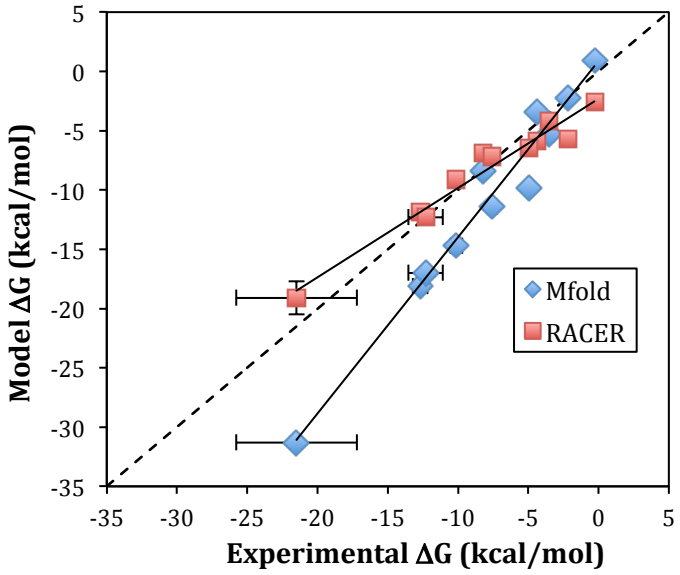


Figure 2.5: Correlation plot between predicted free energy from RACER and experimental free energy in kcal/mol. RACER simulation predicted free energy is compared with Mfold minimum free energy as well as unity slope (dashed line). RACER and Mfold have the same correlation free energy predictive capability (R^2 = 0.96 RACER, R^2 = 0.97 Mfold to experimental free energies), but RACER has a slope closer to unity (slope = 0.75), while Mfold over-stabilizes the free energy of larger RNAs (slope = 1.49). Error bars present on RACER data come from a Monte Carlo bootstrap error analysis as implemented in WHAM by Alan Grossfield[119] (most errors are within the data point).

Pulling generated RNA structures. Ensemble model structures for folded states are shown in Figures 2.S5-2.S6. In the folded states, TAR and h3 are observed to form helices while h1 forms a few base pairs and h2 remains condensed with stacking and base pairs but without helical structure. For duplexes, the two RNA strands form canonical base pairs resulting in proper helices. The terminal nucleotides of d90 are observed to break base pairing with one nucleotide rotating out of the helix while the other remains stacked, but this is also observed in experiment.[126]

In pulling experiments, free energy vs end-to-end extension plots show two distinct energy minima corresponding to folded and unfolded states [127-129]. In the RACER model unfolded (extended) states remain stabilized by vdW_{eff} base stacking interactions, so the location of unfolded free energy is difficult to determine directly from free energy landscapes of RNAs. While the free energy landscapes predicted by RACER show an energy well around the folded state, there is a flat to monotonically increasing curve observed at large extensions (Figures 2.6-2.7, blue curve). To find the unfolded state, we plotted the gradient of the free energy, the 'force' as a function of extension (Figures 2.6-2.7, black curve). From these force vs. extension plots, the predicted free energy of the unfolded state ($\omega_{unfolded}$ from Eq. 5) was taken to be the free energy value where the force is very low (~0.1 kcal/mol/Å), i.e. before the RNA reaches the over-stretched regime (Figures 2.6-2.7, red lines). A 4Å running average of 'force' over extension was used to eliminate noise (Figures 2.6-2.7). Histogram figures showing equal sampling of the pulling windows are included in Figures 2.S7-2.S9. Additionally, the uncertainty of the free energy landscape as computed by a Monte Carlo bootstrap error analysis in the WHAM program by Alan Grossfield[119] is shown as a range in Figures 2.S10-2.S12.

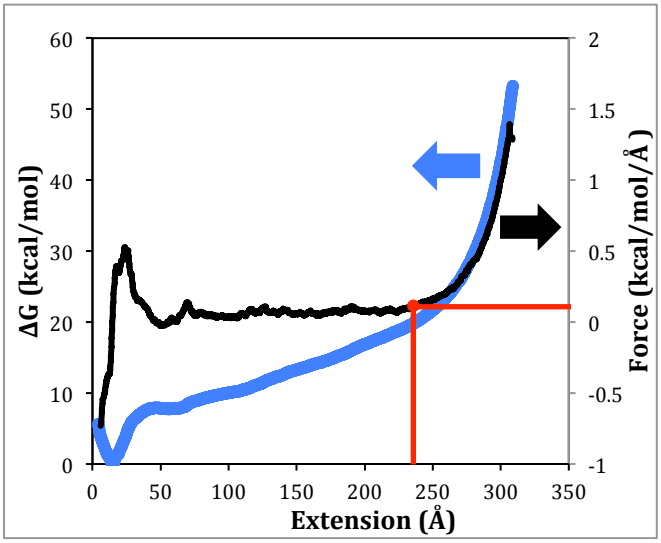


Figure 2.6: The equilibrium pulling free energy profile (blue) of TAR hairpin computed with WHAM using the RACER model (see Method section details). The calculated folding free energy (ΔG) for TAR is -19.1±1.39 kcal/mol. The unfolded state is determined as the state right before the force (derivative of the free energy, curve shown in black) sharply increases from low (< 0.1 kcal/mol/Å) to high due to overstretching. 0.1 kcal/mol/Å and the location of the unfolded state are denoted by the red lines. The experimental value is ≈-21.5±4.3 kcal/mol. A 4Å running average of force (black curve) is shown to eliminate noise.

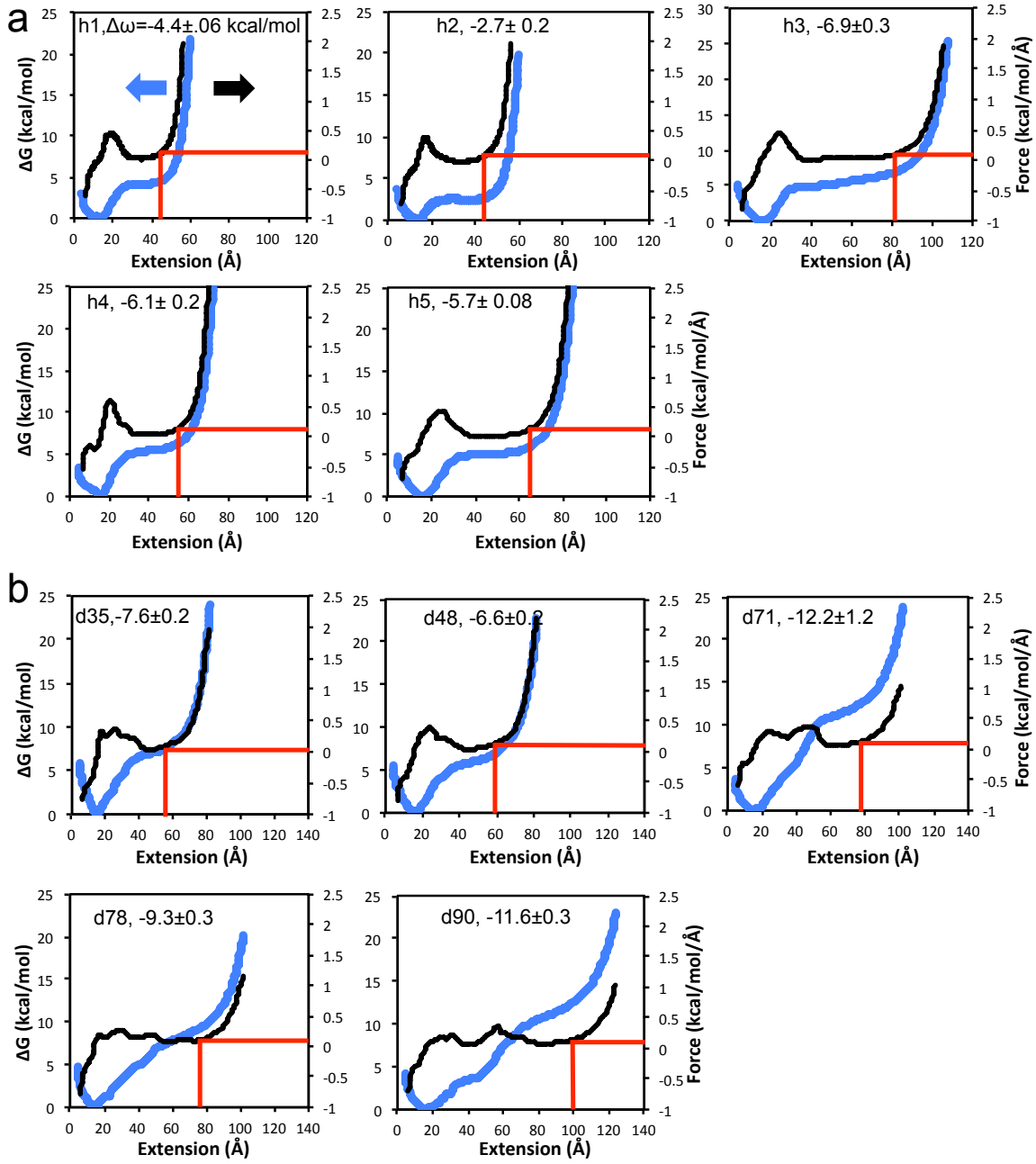


Figure 2.7: The equilibrium pulling free energy profile (blue) of (**a**) hairpins h1-h5 and (**b**) duplexes d35, d48, d71, d78, and d90 computed with WHAM using the RACER model. Umbrella sampling pulling simulations were run for 1μs for each window, with a 1Å window separation. The unfolded state is determined as the state right before the force (derivative of the free energy, curves shown in black) sharply increases from low (< 0.1 kcal/mol/Å) to

high due to overstretching. 0.1 kcal/mol/Å and the location of the unfolded state are denoted by the red lines. The PMF folding free energy (Δω, kcal/mol, not the same as ΔG) is included for each RNA. A 4Å running average of force (black curves) is shown to eliminate noise.

2.3 DISCUSSION

Statistical potential summary. RACER, a coarse-grained RNA model, can accurately predict native structures and capture RNA folding free energy. The functional forms and parameters in RACER were determined by systematic optimization against native structures and melting free energies for a number of RNA molecules. We found that the statistical potentials[26] used in the previous model were over stabilizing and the 3D PMFs diverged at long distances. As a result, we treat RNA as a one-dimensional rather than three-dimensional molecule, and use a 1D RDF when fitting to PMFs. Our optimization procedure led us to incorporate a more general effective van der Waals potential energy function (vdW_{eff}) to describe the interactions among pseudoatoms.

As a result of implementing a new non-bonded potential energy, we have also reparametrized both electrostatic and hydrogen bond potential energy functions. As the RNA backbone is highly charged, a Debye-Huckel electrostatics term is included for each phosphate pseudoatom; a dielectric of 25 was chosen in order to capture both folded and unfolded RNA structures. A directional hydrogen bond potential was reparametrized in order to accurately distinguish base pairing (hydrogen bond, some vdW_{eff}) and base stacking (vdW_{eff}) interactions. We found that the hydrogen bond potential was pivotal to accurate folding free energies as both folded and unfolded RNA have base stacking interactions, while only folded RNA have base pairing (hydrogen bond) interactions.

Thermodynamic summary. For a structure prediction model, thermodynamic accuracy is important to ensure that the energy landscape correctly represents RNAs with

varying size and sequence. Our energy landscape analysis suggests that even relatively small RNAs may have complex energy landscapes, and there are many RNA structures at low potential energy. Therefore, explicit consideration of entropy through techniques such as MD is crucial to capture the free energy landscapes of RNA structures.

Folding free energy values for six RNA hairpins of size 10-52nts and five duplexes of size 6-10bp were determined by umbrella sampling simulations with WHAM-computed free energy. For hairpins, we determined that umbrella sampling simulations with a reaction coordinate of end-to-end extension is appropriate for capturing folding free energy. For duplexes, the same protocol is found to be appropriate, with the addition of a restraint preventing the single strands from long-lasting intra-strand interactions (e.g. hairpin-like structures). Pulling free energy landscapes of hairpins and duplexes clearly showed the folded state and we used the gradient (force) of pulling free energy to define the location of the unfolded state.

Given the low computational cost of RACER, over 0.5 ms of umbrella sampling and simulated annealing simulations are presented. Overall, the MD-calculated free energy results using the RNA model are in excellent agreement (R^2=0.96) with experimental folding free energy values while preserving accurate structure prediction. In this work, we present RACER, a novel RNA coarse-grained model that captures both RNA structure and thermodynamics for increased utility to RNA folding investigations.

2.4 METHODS

Mapping from all-atom to coarse-grained structures. A notable feature of our model is the ability to map to and from all-atom experimental crystal structures. Each of our model's pseudoatoms represents an atomic site in nucleotides; for example, the sugar pseudoatom is assigned the C4' atom position on ribose. Moreover, our model captures

the planarity of the nucleobase with three pseudoatoms. Given a novel (structure undetermined) RNA sequence, our model can first predict the three-dimensional structure in coarse-grained coordinates and then map to all-atom coordinates with further minimization, producing an equivalent to an all-atom experimentally determined structure. As a result, our RNA model is well suited to perform multiscale simulations in the future.

Pulling methods. Melting and pulling experiments are modeled by using umbrella simulations pulling the RNA molecule apart from its terminal ends. A harmonic potential of 1 kcal/mol/Å^2 spring constant is used to restrain the RNA ends at the sugar pseudoatoms (C4' sugar atomic site). Simulation extensions ran from 5.5Å up to fully extended lengths (59.5, 106.5, and 307.5Å for 10, 18, and 52nt hairpins assuming 5.9Å per nt contour length) with a spacing of 1Å between windows.

Duplexes are similarly pulled apart from the sugar pseudoatoms at one terminal end with a 1kcal/mol/Å^2 spring constant; the other terminal end is restrained between two terminal sugar pseudoatoms with a 1 kcal/mol/Å^2 spring constant. Duplex extensions ranged from 5.5 Angstroms up to fully extended lengths (80.5, 100.5, and 124.5 Angstroms for 6, 8, and 10 base pair duplexes respectively) with umbrella window spacing of 1Å. For the duplexes and shorter hairpins of size 10 and 18nt, 1μs of Molecular Dynamics was run for each window. For the TAR hairpin, 100ns was found to be sufficient given the longer end-to-end extension (more windows) needed. We used a 4fs time step for pulling simulations. From the umbrella simulations, the free energy landscapes were computed by the Weighted Histogram Analysis Method[130] (WHAM) using the program distributed by Alan Grossfield[119].

Computational efficiency of the RACER Model. All annealing and pulling simulations (total of 0.5ms) were computed on a local computer cluster. For all

simulations discussed below a 4fs time step was used, and the CPUs used are an early generation Intel Xeon E5345 2.33GHz CPU. Using one CPU core for each simulation, 1μs of simulation of the 10nt hairpin h1 took 22 hours, 1μs of simulation of the 18nt hairpin h3 took ~60 hours, and 100ns of simulation of the 52nt hairpin TAR took ~48 hours. Additionally, 1μs simulation of duplex d35 required 30 hours, while 1μs for duplex d90 required 74 hours. Recently, RACER has been implemented with OpenMP allowing parallelization to multiple cores. In the future, we will implement our model on GPUs, using the software package OpenMM[131]. Implementation of RACER on GPUs will allow for even better efficiency. As a result of the improved computational efficiency offered by the coarse-graining, it will be possible to simulate RNAs at physiologically relevant timescales.

Implementation and parameters. The TINKERMD implemented RACER model is available free of charge at http://biomol.bme.utexas.edu/tinker-openmm/index.php/TINKER-OPENMM:Development-rna. The parameters and conversion programs are included in the distribution. Conversion tutorials are posted online at http://biomol.bme.utexas.edu/tinker-openmm/index.php/TINKER-OPENMM:Tutorials-rna.

2.5 ADDITIONAL FIGURES AND DISCUSSION

$$E_{vdW\_eff} = \frac{2\varepsilon}{1-\frac{3}{\gamma+3}}\left(\frac{\sigma^6}{\sigma^6+r^6}\right)\left[\frac{3}{\gamma+3}e^{\gamma\left(1-\frac{r}{\sigma}\right)}-1\right]$$

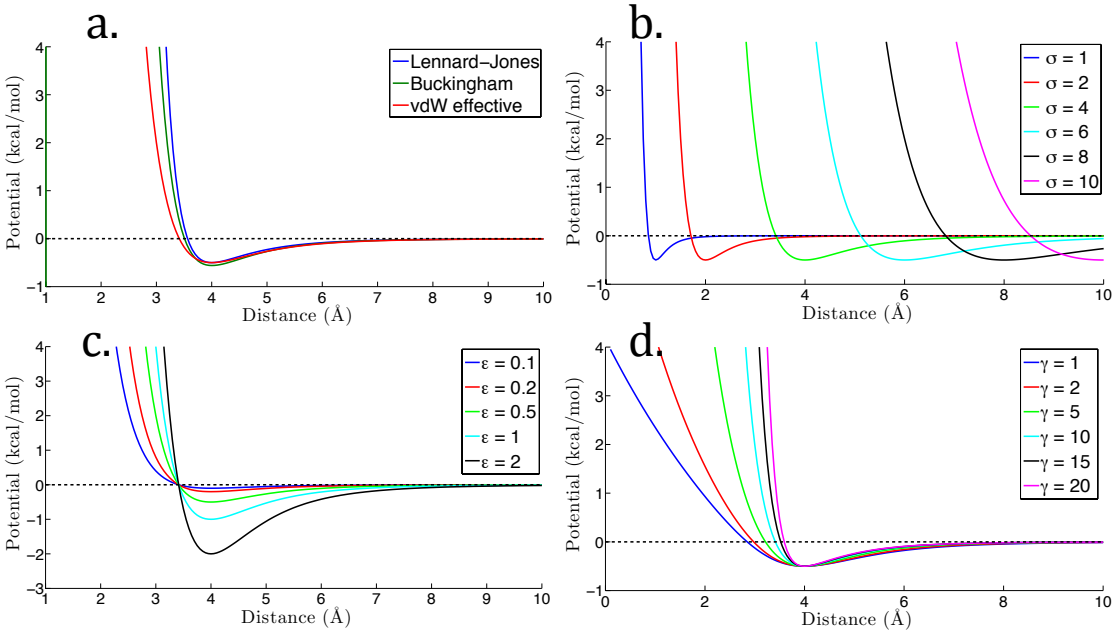
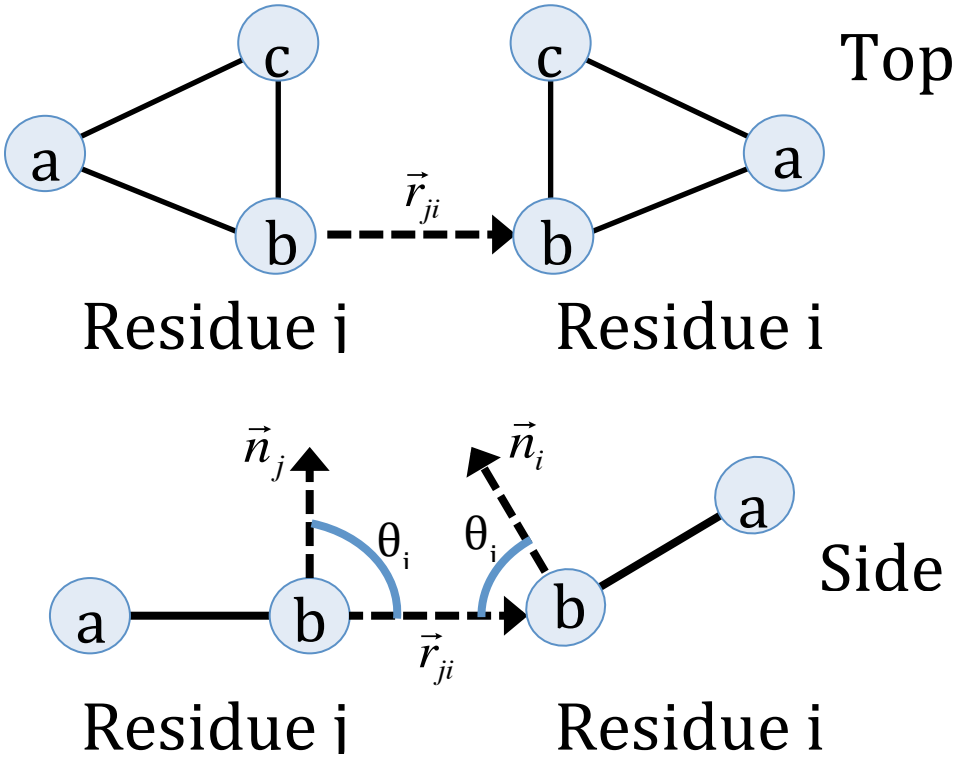


Figure 2.S1: vdW_{eff} potential. (a.) Effective potential compared to standard Lennard Jones and Buckingham potentials with minimum energy potential ε = 0.5 kcal/mol, minimum energy distance, σ = 4Å, and gamma of effective potential γ = 10. (b.) Effect of changing value of minimum energy distance, σ (c.) Effect of changing minimum energy potential, ε (d.) Effect of changing the short range behavior with parameter γ. For (b-d), unless stated, ε = 0.5 kcal/mol, σ = 4Å, and γ = 10.



$$E_{hb} = -\frac{\varepsilon_{hb,max}}{2}(1-\cos(\alpha_k))\left(\frac{\sigma_{hb,eq}}{|\overrightarrow{r_{ji}}|}\right)^3$$

$$\alpha_k = 2(\theta_i+\theta_j)-\pi, \qquad \frac{\pi}{2} < (\theta_i+\theta_j) < \frac{3\pi}{2}$$

$$\theta_j = \cos^{-1}\left(\frac{\overrightarrow{n_j}\cdot\overrightarrow{r_{ji}}}{|\overrightarrow{n_j}||\overrightarrow{r_{ji}}|}\right), \qquad \overrightarrow{n_j} = \overrightarrow{r_{jab}}\times\overrightarrow{r_{jcb}}$$

Figure 2.S2: Hydrogen bond potential diagram and equations. $\overrightarrow{n_i}$ and $\overrightarrow{n_j}$ are the vectors normal to the plane of residues i and j respectively. $\overrightarrow{r_{jab}}$ is the vector from atom b to atom a on residue j and $\overrightarrow{r_{jcb}}$ is the vector from atom c to atom a on residue j. $\overrightarrow{r_{ji}}$ is the vector between hydrogen bonding atoms of residues j and i. θ_i and θ_j are the angles between the respective normal vectors and vector $\overrightarrow{r_{ji}}$.

Table 2.S1: Simulated annealing energy landscapes for 14 PDB structures. PDB ID is stated at the top of each plot. The total potential energy as a function of RMSD to the PDB structure are shown. For annealing protocol, see main text.

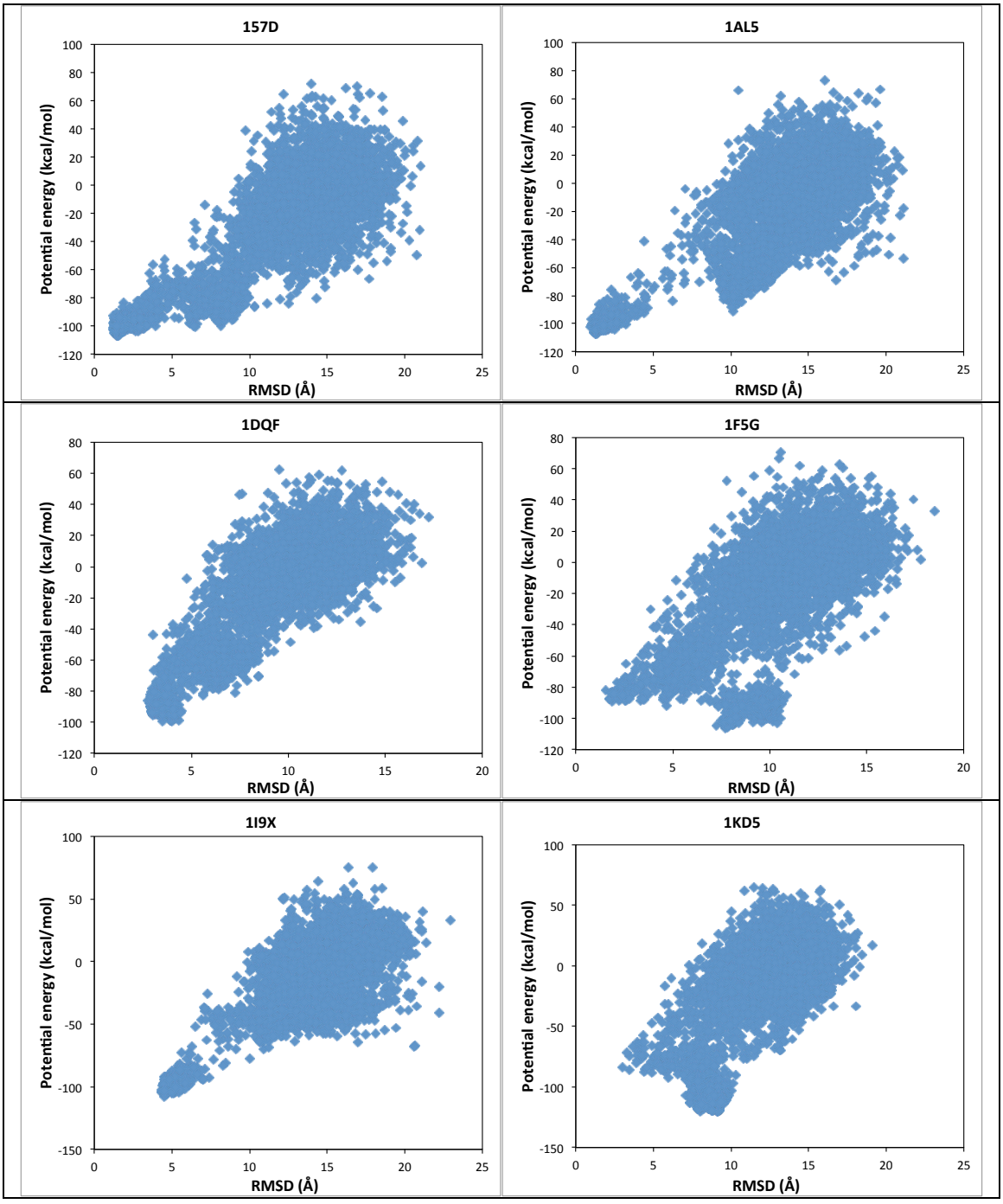


Table 2.S1 (continued)

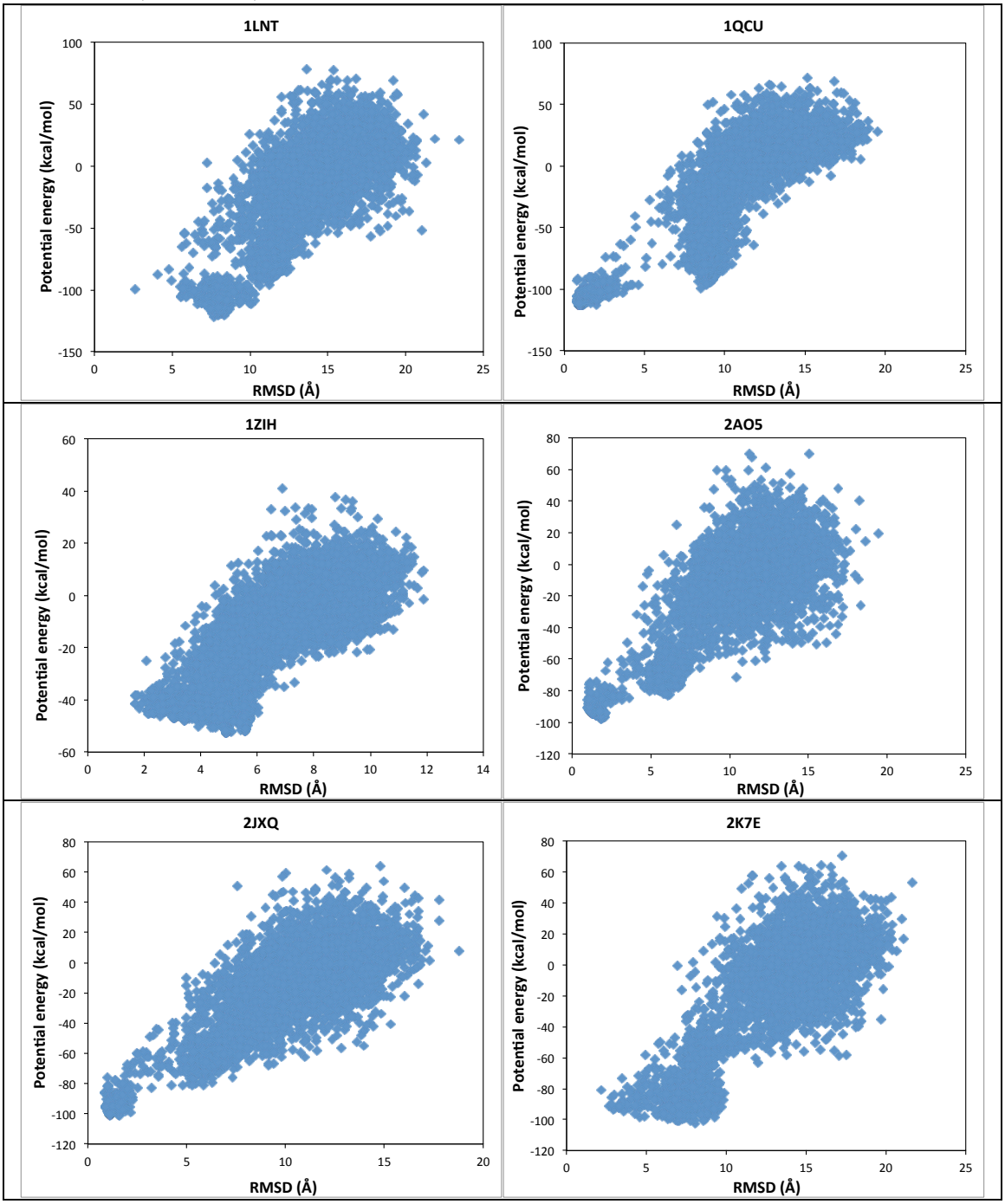
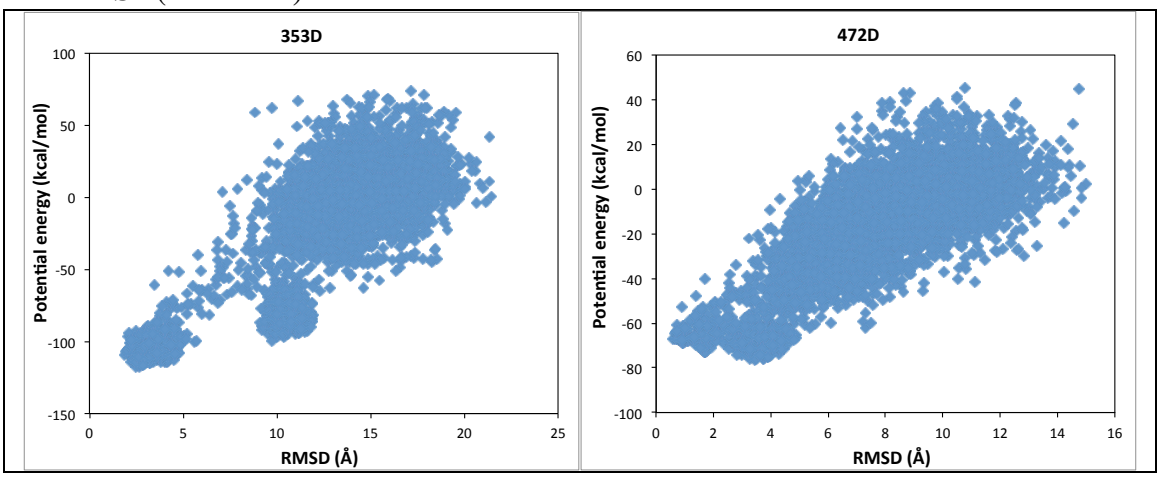


Table 2.S1 (continued)



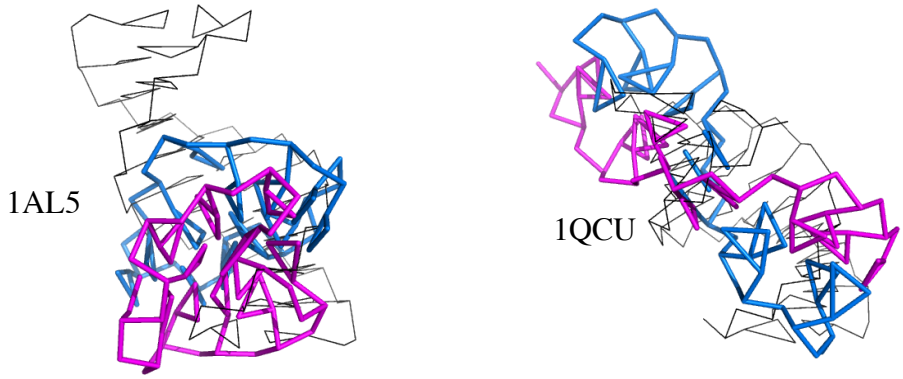


Figure 2.S3: Annealing structures taken from bottom of funnel feature: 1AL5 (left) and 1QCU (right). Annealing structures are colored blue and magenta while experimental structures are colored black. Note the extended, base-stacking structure of 1QCU.

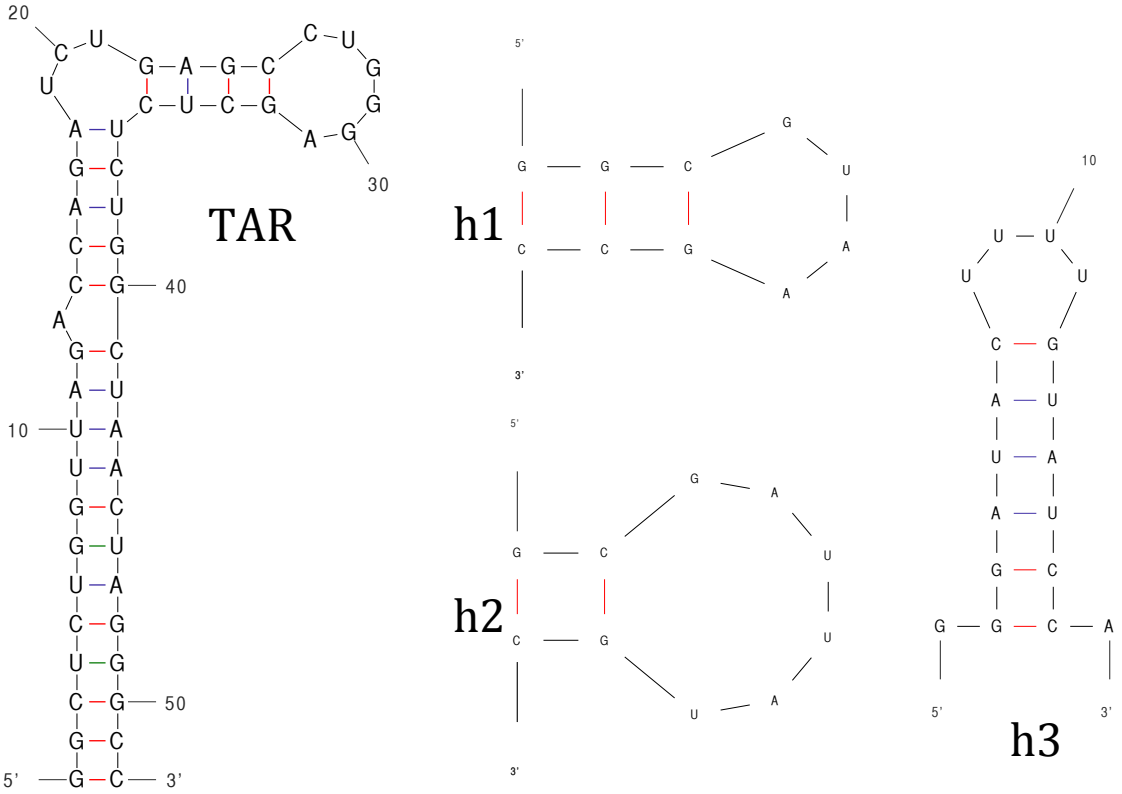


Figure 2.S4: Mfold predicted minimum free energy secondary structures for the hairpins reported: TAR (left), and Turner hairpin sequences h1 (top, middle), h2 (bottom, middle), and h3 (right).

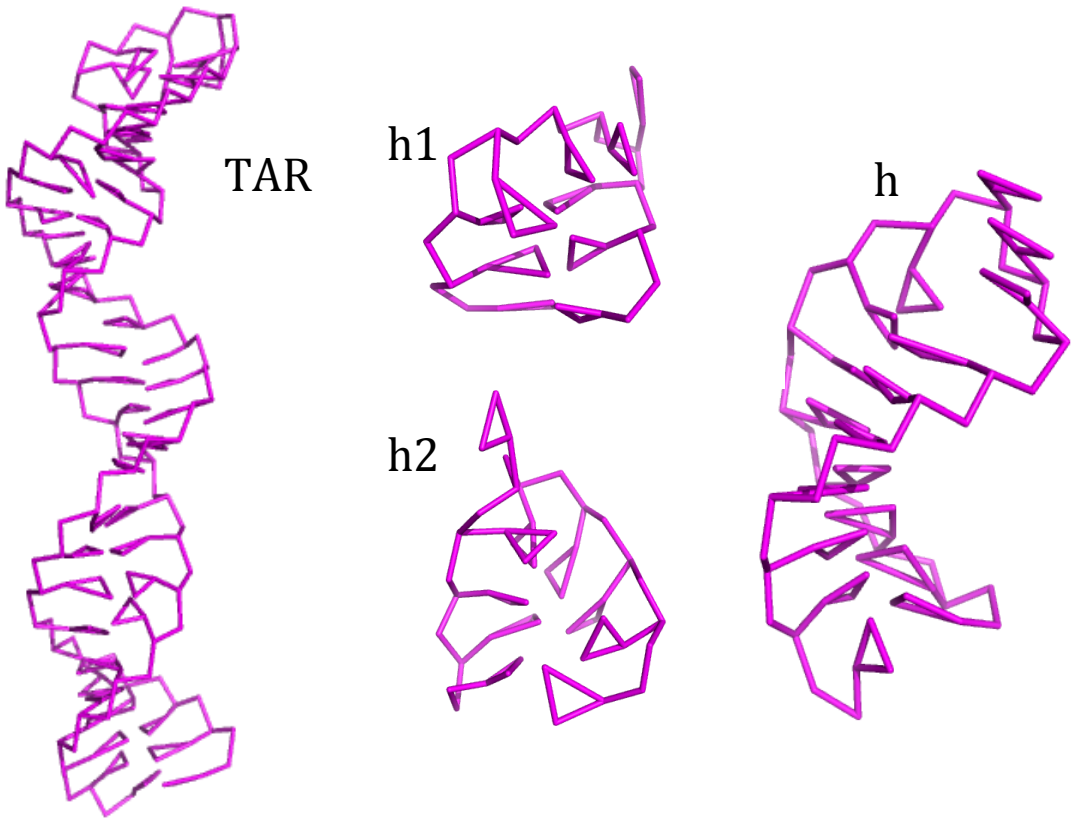


Figure 2.S5: Model structures of hairpins used for pulling sequences. Structures are taken from the ensemble of equilibrium end-end extension structures.

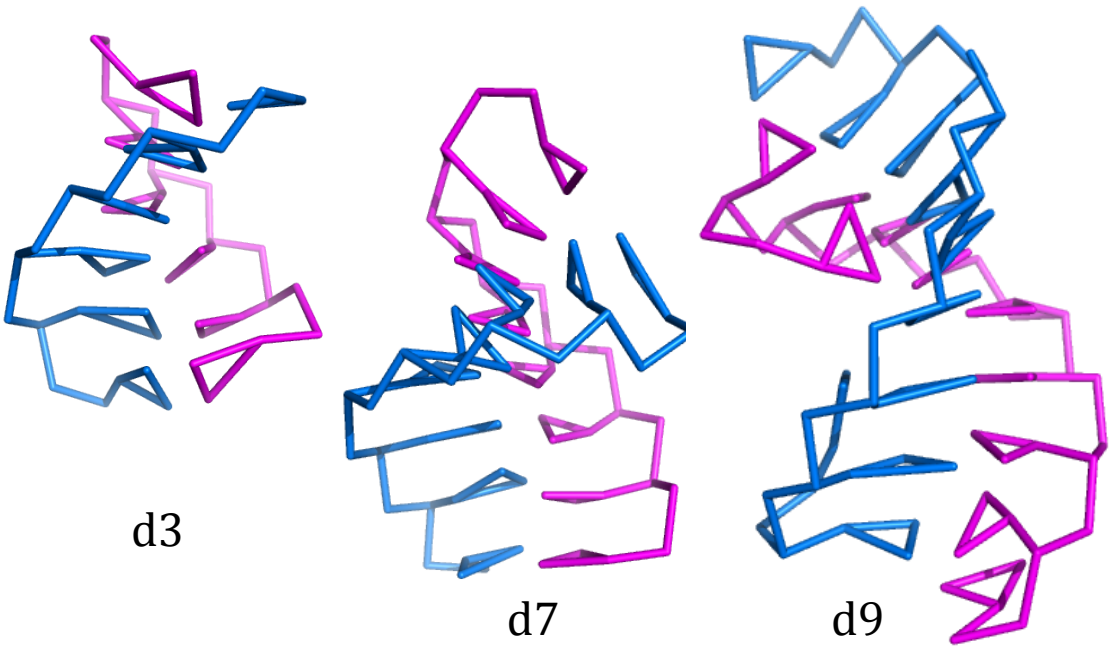


Figure 2.S6: Model structures of duplexes used for pulling simulations. Structures are taken from the ensemble of equilibrium structures.

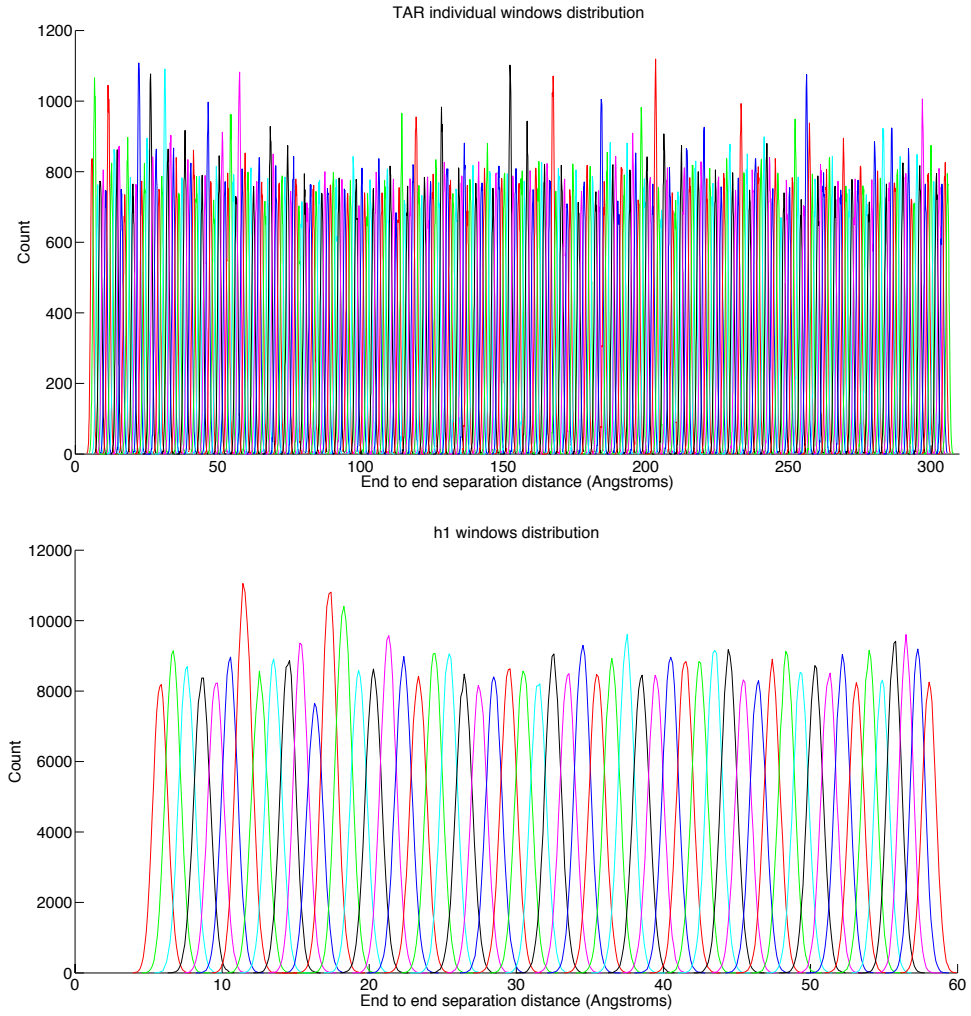


Figure 2.S7: Sampling distribution of each umbrella sampling window for both TAR (top) and h1 (bottom). The separation distance between windows was 1Å for all RNAs.

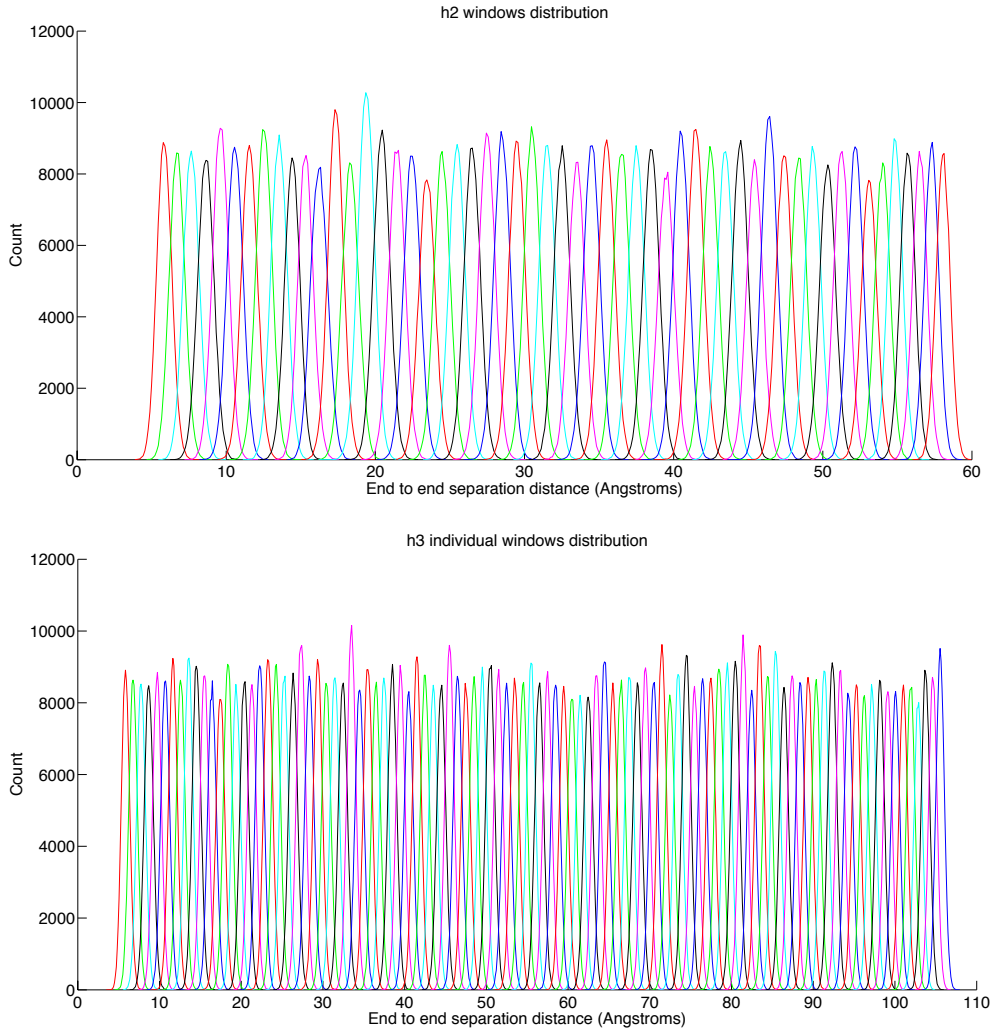


Figure 2.S8: Sampling distribution of each umbrella sampling window for both h2 (top) and h3 (bottom). The separation between windows is 1Å.

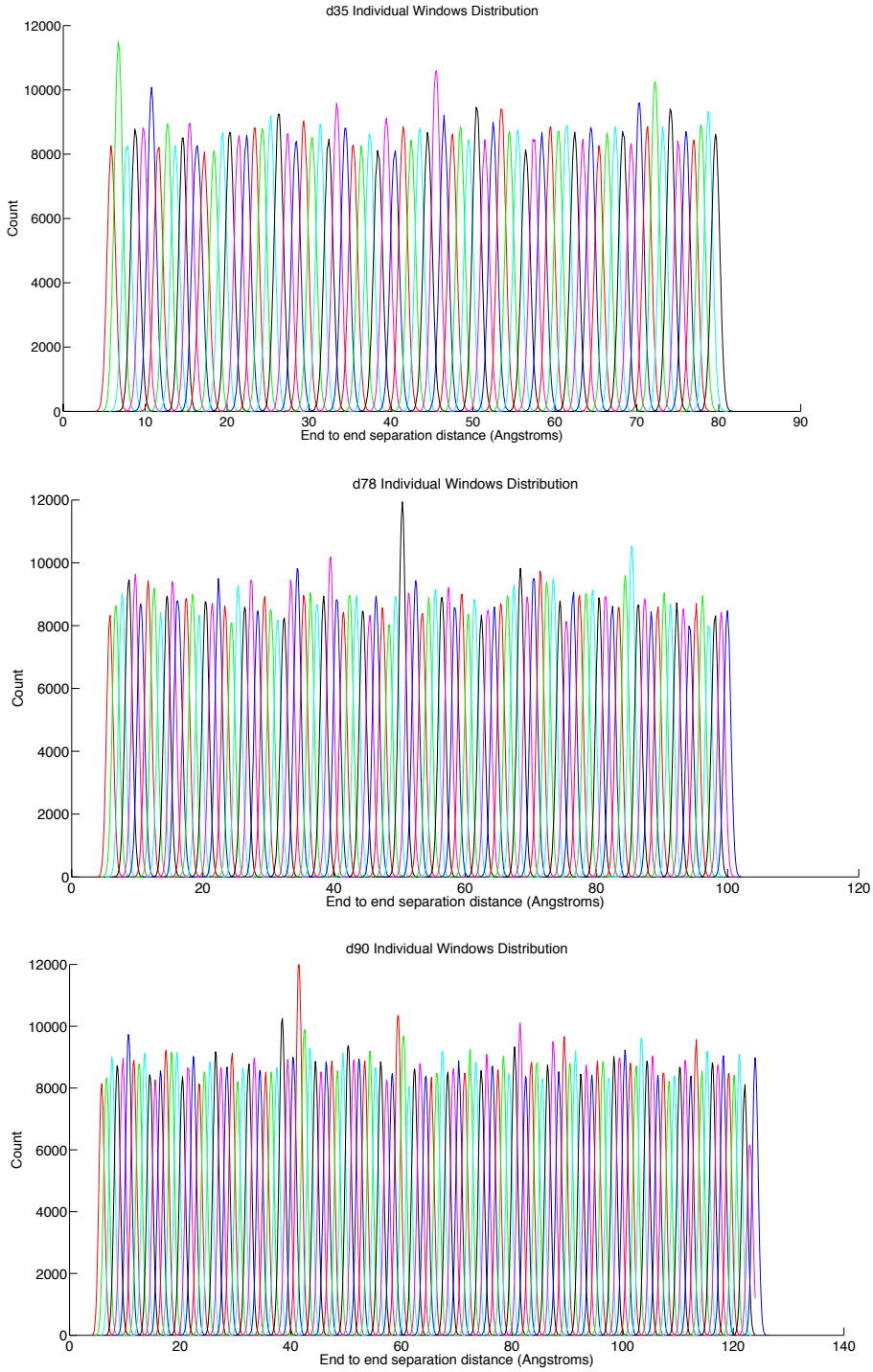


Figure 2.S9: Sampling distribution of each umbrella sampling window for d35 (top), d78 (middle), and d90 (bottom). The separation between windows is 1Å.

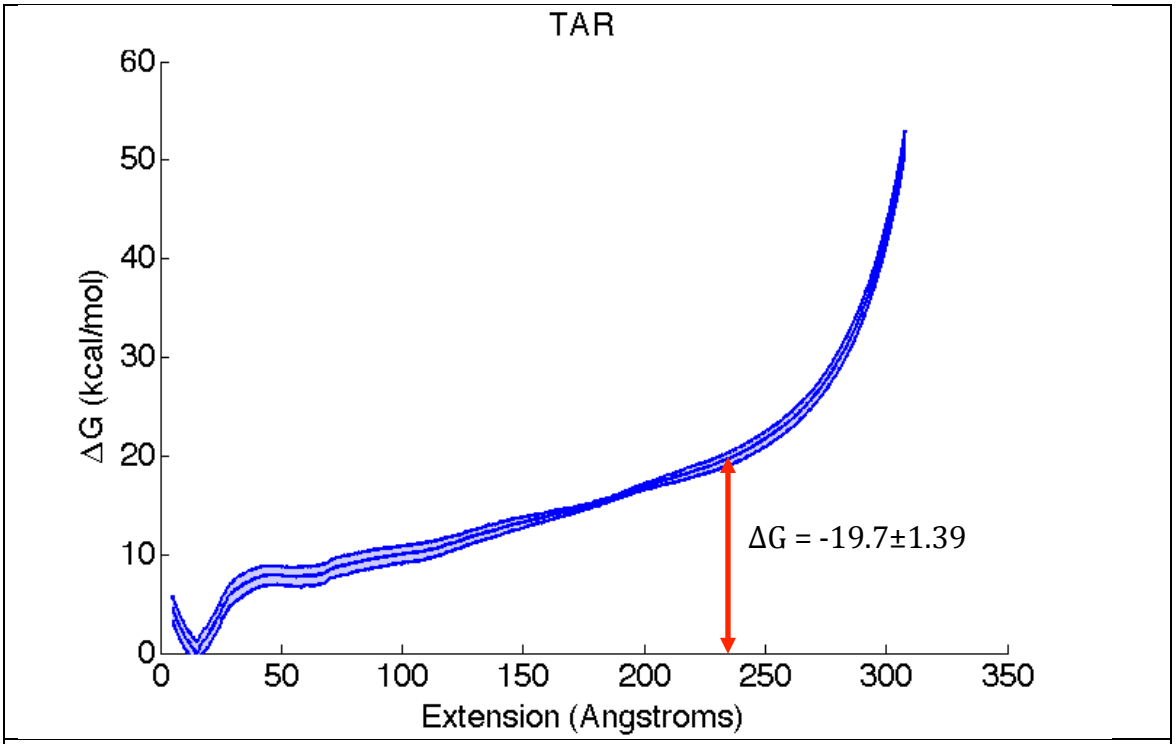


Figure 2.S10: TAR pulling free energy landscape with computed error shown as range. Error is take from a Monte Carlo bootstrap error analysis as implemented in the WHAM program by Grossfield[119]. The RACER predicted free energy is -19.7±1.39 kcal/mol; the experimental value is ≈-21.5 kcal/mol.

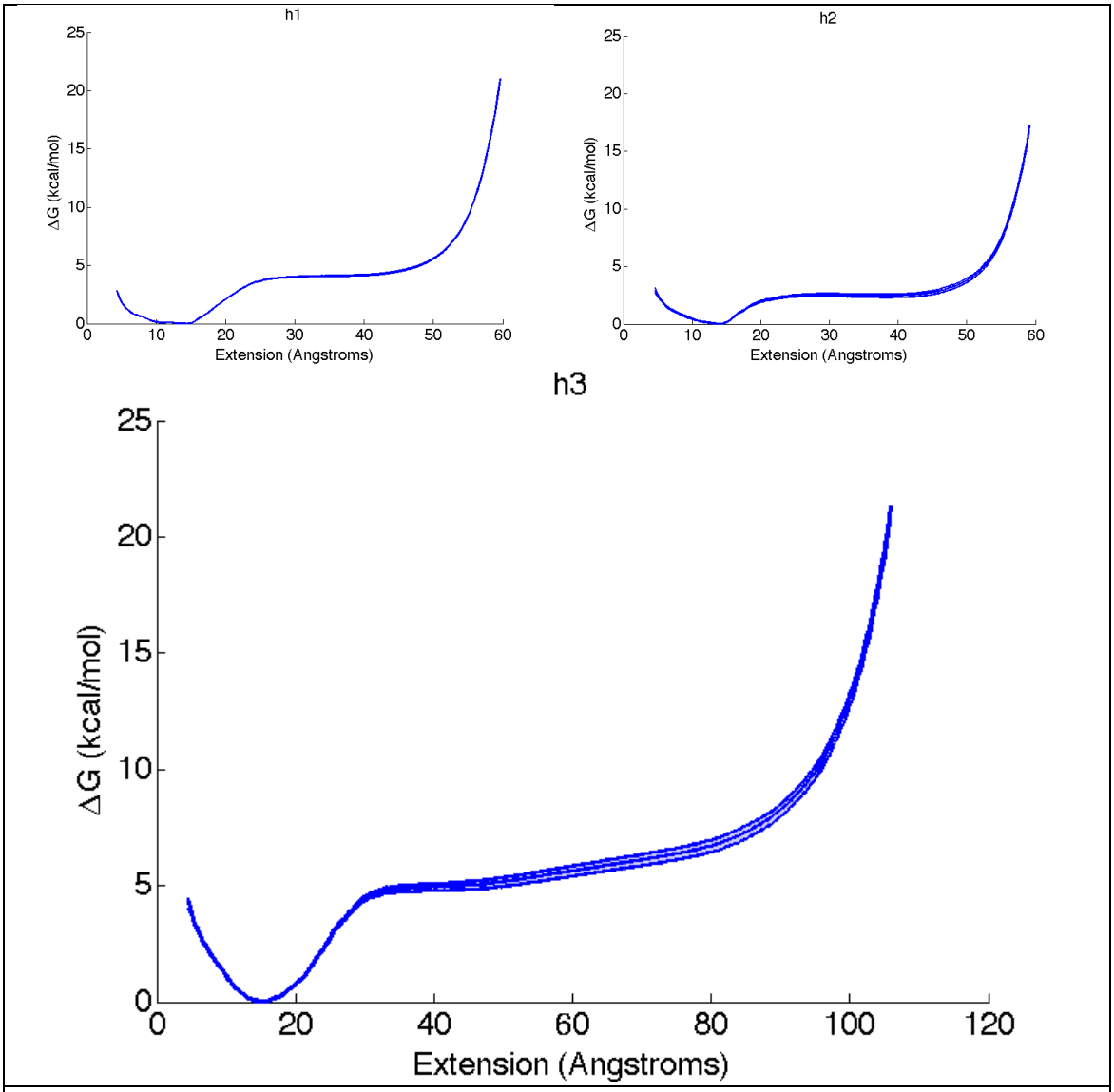


Figure 2.S11: Hairpin h1(top, left), h2 (top, right), and h3 (bottom) pulling free energy landscapes with computed error shown as range. Error is take from a Monte Carlo bootstrap error analysis as implemented in the WHAM program by Grossfield[119].

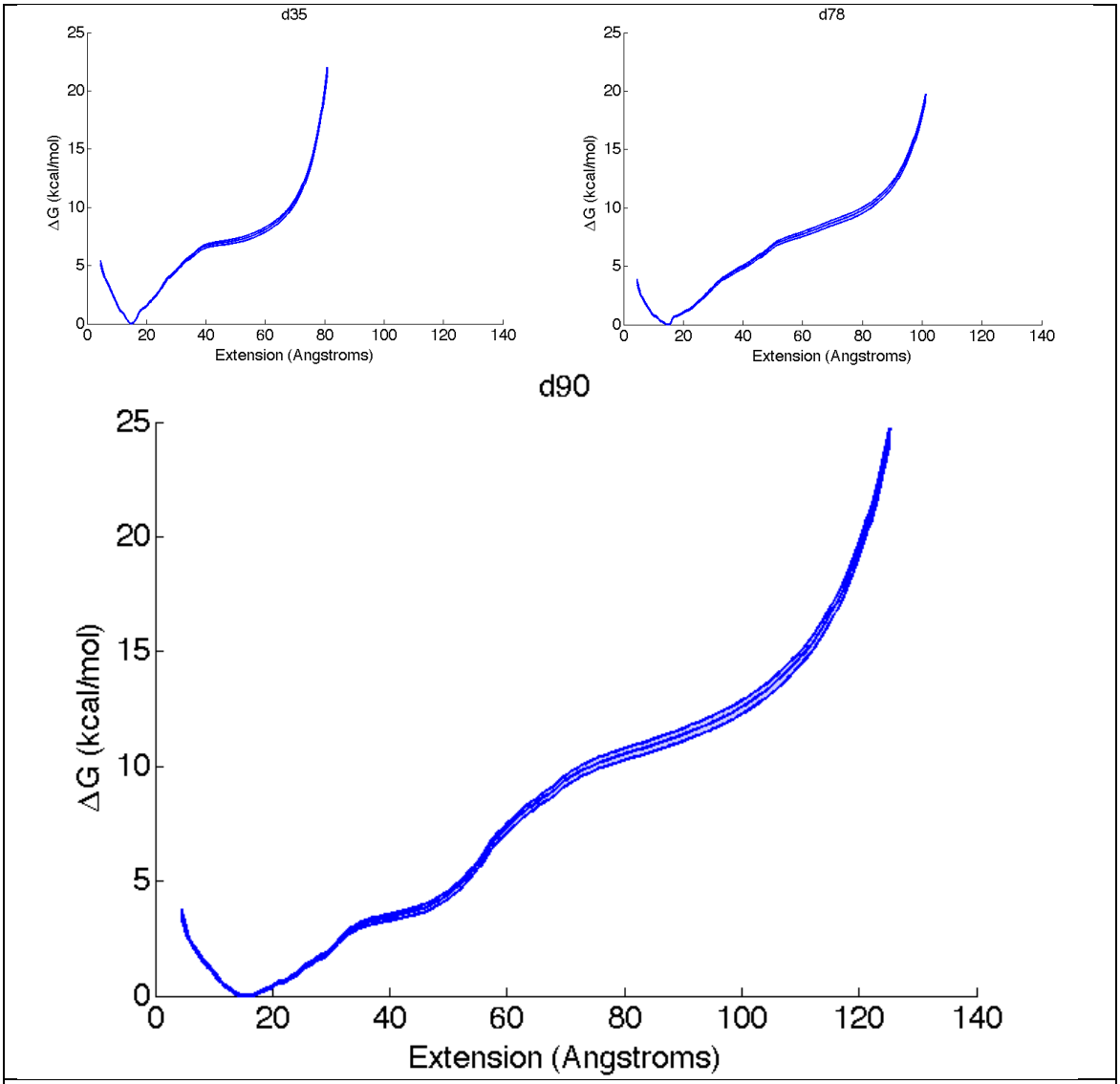


Figure 2.S12: Duplex d35(top, left), d78 (top, right), and d90 (bottom) pulling free energy landscapes with computed error shown as range. Error is take from a Monte Carlo bootstrap error analysis as implemented in the WHAM program by Grossfield[119].

Debye-Huckel parameterization. The electrostatics potential also contributes to non-bonded interactions and should be included in fitting to the PMF curves. We used the same Debye-Huckel electrostatics term shown in Eq. 4 for our model. In RNA nucleotides, only the phosphate group is net-charged and correspondingly, only the phosphate pseudoatom in our model has a charge of -1.0*e*. The Debye-Huckel potential is purely repulsive, so this must be balanced by attractive terms with the vdW_{eff} potential, when fitting to the phosphate nonbonded PMF. Note that this potential incorporates an implicit solvent effect in addition to Coulomb interactions. During optimization, the Debye-Huckel potential was fit concomitantly to the vdW_{eff} potential for both structure and energy. In order to select an optimal dielectric constant and Debye-length for our model, we analyzed the effects of varying the dielectric constant and Debye-length on the RMSD and Pearson R^2 correlation of our model RMSD and experimental energies for 14 PDB structures listed in Table 1. As seen in Fig. S13, our selection of a dielectric constant of 25 and a Debye length of 10 Angstroms corresponds to both low RMSD and high R^2 correlation. Typical dielectric constants range from ~78 for pure water[132] to <5 for folded proteins[133,134]. Small RNAs are much more exposed than compact proteins. Though the dielectric constant will vary over RNA structure, a dielectric value of 25 is a reasonable compromise between folded and unfolded states. Whereas the dielectric is a proportionality constant, the Debye-length ξ acts as a decay constant, describing how quickly the electrostatic energy decreases over distance. If we assume a monovalent salt

concentration of 0.1 M, the Debye-length is approximately 9.6 Angstroms[135], so the Debye-length of 10 Angstroms used here is appropriate.

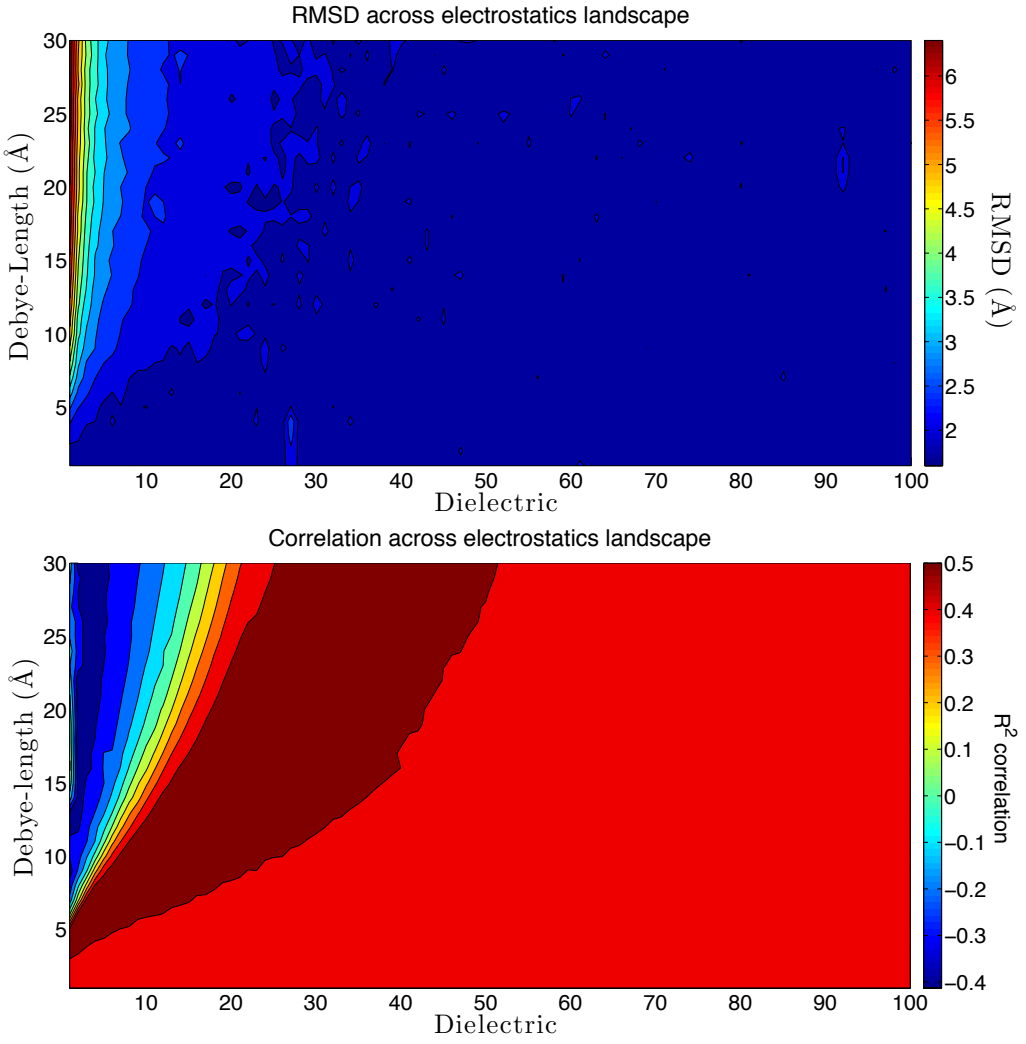


Figure 2.S13: RMSD (top) and Pearson R^2 correlation coefficient (bottom) values as a function of Dielectric and Debye-Length (Å). The RMSD value is the average of 14 PDB structures averaged over 5ps molecular dynamics simulation. Pearson product moment correlation coefficient (R^2) is between RACER model potential energy after minimization and experimental melting free energies for a set of 90 RNA sequences taken from ref[112].

Torsion parameterization. Parameterization of model torsion interaction required consideration of double counting (overlap between the nonbonded and torsional terms) and formation of non A-form helices. Similar to nonbonded interactions, when torsion parameters were naively fit to the PDB structure-statistics derived PMF, the torsion potential was excessively large (3x bending and angle potential energies). To better understand this, we studied the residue-level torsion interactions and determined that for the raw-PMF torsion parameters, torsion interactions are at least double and in some cases quadruple counted since multiple torsional terms are describing the exactly same rotation. For instance, in the backbone S-P' pseudoatom connection, there exists both torsion terms PSP'S' as well as C*SP'S', where prime(') represents the adjacent nucleotide and C* represents the pseudoatom type CU or CG. Although these torsion terms depict distinct chemical structures, they are not independent. Further, the S-C* pseudoatom connection is quadruple counted, with interactions PSC*B1, PSC*B2, B1C*SP', and B2C*SP' where B1 and B2 represent two separate base atoms. A correlation figure of 2 S-C* torsions is shown in Figure S14. Note that if the S-C* torsions were not correlated, Figure S14 would be a uniform blue rectangle; the concise sampled area is a clear indication of correlation. To address the above issue, we reduced the k_n torsional force constants to half (or quarter in some cases) of their PMF-fit values. In addition, we noticed that our model tends to over predict A-form helices. This is likely due to the strong biases intrinsic to the RNA structural statistics. We therefore expanded the range of allowed torsion angles beyond the statistical PMF. Another important

consideration is that the Boltzmann inversion from structural-statistics assumes torsion or pair-wise PMF is independent of all other interactions. What the "torsion" PMF captures is a combination of atomic repulsion-dispersion and electrostatic interactions. In our model vdW_{eff} interactions between 1-4 bonded pseudoatoms are not computed. However, we still found considerable interplay between torsion and vdW_{eff} interactions, and the balance of these terms must be carefully accounted for in the parameterization of effective nonbonded potential. Ultimately, by consideration of double counting and structural flexibility, we optimized the torsion potential to adequately depict intra-strand rotation.

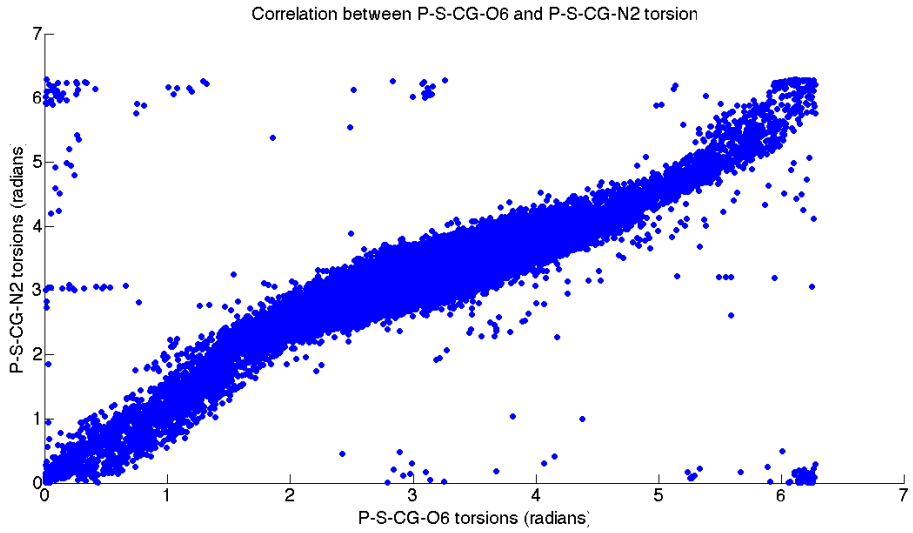


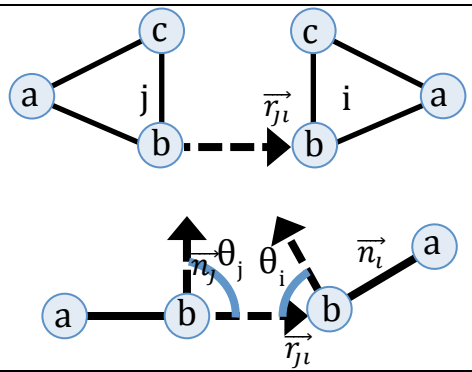
Figure 2.S14: Correlation of torsions P-S-CG-O6 and P-S-CG-N2 from Guanosine nucleotides. Over 70,000 PDB Guanosine nucleotides were sampled; uncorrelated torsions would yield a completely uniform blue figure. The concise region of blue samples indicates correlation between these two torsion angles.

Hydrogen bond details. Hydrogen bond potential (Eq. 3) was kept in our model from our previous work[100]; however, with the introduction of the new vdW_{eff} potential, we re-optimized hydrogen bond parameters to distinguish between base pairing and base stacking interactions. For our model, equilibrium hydrogen bond length $\sigma_{hb,eq}$ is taken to be 2.9Å while $\varepsilon_{hb,max}$ is taken to be 2.0 kcal/mol. Table 2.S2 presents representative potential components for stacking and base pairing interactions taken from PDB ID: 1AL5 structure (A-form dodecamer[136]). Using only the vdW_{eff} potential term, it is observed that base stacking is stronger by ~0.5 kcal/mol over base-pairing interactions. However, with the addition of the hydrogen bond term, base pairing becomes slightly stronger. The GC base pair potential is observed to be 2 kcal/mol more stabilizing than AU base pair largely as a result of hydrogen bonding. This mimics the canonical interactions of the GC basepair with three hydrogen bonds while the AU basepair has two hydrogen bonds. The hydrogen bond term acts on every base pair and is therefore semi-proportional to the number of base pairs. The proportionality helped to increase our predicted folding free energy values to agree better with experiment. Additionally, direct proportionality of the potential energy to base pairs is not desirable since GC, AU, as well as non-canonical basepairs have different stabilities; this leaves room for vdW_{eff} interactions to provide atom type based discrimination.

Table 2.S2: Potential values for representative stacking and canonical base pair structures taken from PDB ID: 1AL5.

	Potential (kcal/mol)		
Structure	vdW_{eff}	Hbond	Total
AU Stack	-2.45	-	-2.45
GC Stack	-3.04	-	-3.04
AU base pair	-2.01	-1.21	-3.22
GC base pair	-2.41	-2.89	-5.30

Directionality and derivative of hydrogen bond potential. Another interesting aspect of the hydrogen bond potential implemented here is that it is directional without including a hydrogen atom, unlike most hydrogen bond potentials [137]. The potential used here accounts for interaction distances and angles of the base pairs by taking the angle between the normal vectors of the bases (see Fig. S15). The analytical force (negative of gradient) of this potential energy, which is needed for energy minimization and molecular dynamics, is given in Fig. S15; notably, the force is dependent on the angle and distance between the bases, with the angle components dependent on positions of all six pseudoatoms.



$$\frac{\partial E}{\partial x_{ia}} = -\frac{\varepsilon_{hb,max}}{2}\left[\sin\left(\alpha_k(x_{ia})\right)\left(\frac{\sigma_{hb,eq}}{\left|\overrightarrow{r_{jl}}\right|}\right)^3\left(\frac{\partial \alpha_k}{\partial x_{ia}}\right)\right]$$

$$\frac{\partial \alpha_k}{\partial x_{ia}} = 2\frac{\partial \theta_i}{\partial x_{ia}} = \frac{-2}{\sqrt{1-\left(\frac{\overrightarrow{n_l}\cdot\overrightarrow{r_{jl}}}{|\overrightarrow{n_l}||\overrightarrow{r_{jl}}|}\right)^2}}\frac{\partial\left(\frac{\overrightarrow{n_l}\cdot\overrightarrow{r_{jl}}}{|\overrightarrow{n_l}||\overrightarrow{r_{jl}}|}\right)}{\partial x_{ia}}$$

$$\frac{\partial\left(\frac{\overrightarrow{n_l}\cdot\overrightarrow{r_{jl}}}{|\overrightarrow{n_l}||\overrightarrow{r_{jl}}|}\right)}{\partial x_{ia}} = \left[\frac{-z_{icb}y_{ji}+y_{icb}z_{ji}}{|\overrightarrow{n_l}|\left|\overrightarrow{r_{jl}}\right|} - \frac{(\overrightarrow{n_l}\cdot\overrightarrow{r_{jl}})(-x_{iab}z_{icb}+x_{icb}z_{iab}-z_{icb}+y_{icb}(x_{iab}y_{icb}-x_{icb}y_{iab}))}{|\overrightarrow{n_l}|^3\left|\overrightarrow{r_{jl}}\right|}\right]$$

$$\frac{\partial E}{\partial x_{ib}} = -\frac{\varepsilon_{hb,max}}{2}\left[\sin(\alpha_k)\left(\frac{\sigma_{hb,eq}}{\left|\overrightarrow{r_{jl}}\right|}\right)^3\left(\frac{\partial \alpha_k}{\partial x_{ib}}\right) - 3(1-\cos(\alpha_k))\left(\frac{\sigma_{hb,eq}{}^3}{\left|\overrightarrow{r_{jl}}\right|^4}\right)\left(\frac{\partial\left|\overrightarrow{r_{jl}}\right|}{\partial x_{ib}}\right)\right]$$

$$\frac{\partial \alpha_k}{\partial x_{ib}} = 2\frac{\partial \theta_i}{\partial x_{ib}} + 2\frac{\partial \theta_j}{\partial x_{ib}}\ , \qquad \frac{\partial\left|\overrightarrow{r_{jl}}\right|}{\partial x_{ib}} = \frac{x_{ib}-x_{jb}}{\left|\overrightarrow{r_{jl}}\right|}$$

$$\frac{\partial \theta_i}{\partial x_{ib}} = \frac{-1}{\sqrt{1-\left(\frac{\overrightarrow{n_l}\cdot\overrightarrow{r_{jl}}}{|\overrightarrow{n_l}||\overrightarrow{r_{jl}}|}\right)^2}}\frac{\partial\left(\frac{\overrightarrow{n_l}\cdot\overrightarrow{r_{jl}}}{|\overrightarrow{n_l}||\overrightarrow{r_{jl}}|}\right)}{\partial x_{ib}}, \quad \frac{\partial \theta_j}{\partial x_{ib}} = \frac{-1}{\sqrt{1-\left(\frac{\overrightarrow{n_j}\cdot\overrightarrow{r_{jl}}}{|\overrightarrow{n_j}||\overrightarrow{r_{jl}}|}\right)^2}}\frac{\partial\left(\frac{\overrightarrow{n_j}\cdot\overrightarrow{r_{jl}}}{|\overrightarrow{n_j}||\overrightarrow{r_{jl}}|}\right)}{\partial x_{ib}}$$

Figure 2.S15 (continued)

$$\frac{\partial\left(\frac{\overrightarrow{n_i}\cdot\overrightarrow{r_{ji}}}{|\overrightarrow{n_i}||\overrightarrow{r_{ji}}|}\right)}{\partial x_{ib}} = \left[\begin{array}{c} \frac{\left((y_{iab}z_{icb} - y_{icb}z_{iab}) + (z_{icb} - z_{iab})y_{ji} + (y_{iab} - y_{icb})z_{ji}\right)}{|\overrightarrow{n_i}||\overrightarrow{r_{ji}}|} \\ -\frac{(\overrightarrow{n_i}\cdot\overrightarrow{r_{ji}})\left((-x_{iab}z_{icb} + x_{icb}z_{iab})(z_{icb} - z_{iab}) + (x_{iab}y_{icb} - x_{icb}y_{iab})(y_{iab} - y_{icb})\right)}{|\overrightarrow{n_i}|^3|\overrightarrow{r_{ji}}|} - \frac{(\overrightarrow{n_i}\cdot\overrightarrow{r_{ji}})x_{ji}}{|\overrightarrow{n_i}|\left|\overrightarrow{r_{ji}}\right|^3} \end{array}\right]$$

$$\frac{\partial\left(\frac{\overrightarrow{n_j}\cdot\overrightarrow{r_{ji}}}{|\overrightarrow{n_j}||\overrightarrow{r_{ji}}|}\right)}{\partial x_{ib}} = \frac{y_{jab}z_{jcb} - y_{jcb}z_{jab}}{\left|\overrightarrow{n_j}\right|\left|\overrightarrow{r_{ji}}\right|} - \frac{(\overrightarrow{n_j}\cdot\overrightarrow{r_{ji}})x_{ji}}{\left|\overrightarrow{n_j}\right|\left|\overrightarrow{r_{ji}}\right|^3}$$

Figure 2.S15: Hydrogen bond potential diagram and derivative (negative of force) equations for the *x* coordinate of atoms *a* and *b* on residue *i*. $\overrightarrow{n_i}$ and $\overrightarrow{n_j}$ are the vectors normal to the plane of residues i and j respectively. $\overrightarrow{r_{ji}}$ is the vector between hydrogen bonding atoms from residues j to i (jb and ib in this case). θ_i and θ_j are the angles between the respective normal vectors and vector $\overrightarrow{r_{ji}}$. x_{ia} is the x-coordinate of atom a on residue i. x_{iab} is the x-coordinate term of the vector from b to a. x_{ji} is the x-coordinate term of the vector $\overrightarrow{r_{ji}}$. For derivatives, atom c follows similarly to atom a.

Chapter 3: Calculating binding free energies of host-guest systems using AMOEBA polarizable force field[2]

3.1 INTRODUCTION

Molecular recognition is fundamental to biological processes and is utilized in applications ranging from therapeutics to chemical sensors[138]. Understanding the importance of molecular recognition, the interactions involved are exceedingly complex and dependent upon a high degree of order between the solutes and the solvent for binding. Computer prediction of binding affinity holds potential to accurately capture thermodynamic information from different states as well as allow for the design of novel ligands.

Molecular modeling and simulation can be a powerful tool for quantitative understanding of the driving forces underlying molecular recognition[139,140]. However, accurate computation of binding free energy via molecular modeling faces two major challenges. First, the energetic description of binding requires high accuracy potential energy that is also transferable between different chemical and physical environments. Second, the flexibility of guest, water and host molecules results in many degrees of freedom making it difficult to adequately explore the configuration space using molecular dynamics. With increasing complexity up to protein-ligand systems, sufficient sampling of binding by traditional methods becomes limited by computational cost.

Numerous potential energy methods have been proposed to compute binding free energy, increasing in complexity from empirical docking methods to quantum mechanics (QM) calculations[141]. Empirical docking methods[142] are frequently used for library screening and though they allow for fast calculation, they do not maintain high accuracy

[2]Large portions of this chapter are based on the work: Bell, DR., et al. Calculating binding free energies of host-guest systems using AMOEBA polarizable force field. *Phys Chem Chem Phys*. 2016

of the potential energy function nor do they allow for sufficient sampling of binding conformations. QM calculations of binding free energy[143-145] are limited to small, predetermined binding sites. Bridging the gap between docking methods and QM, semi-empirical force-field methods use Molecular Dynamics or Monte Carlo sampling schemes to generate many configurations and energies[146,147]. In semi-empirical force field methods, the potential energy of the system is computed from analytical functions of the atomic coordinates. Classical force fields such as AMBER[148], CHARMM[149], OPLS-AA[150], or GROMOS[151] typically represent intermolecular interactions by a van der Waals (vdW) term and point charge electrostatics. This representation is computationally efficient and sufficiently accurate for many applications. However, the potential energy is limited by not capturing electrostatic response to environmental stimulus, referred to as the polarization effect[152,153]. Additionally, modeling electrostatics as point charges neglects the intricate yet substantial effect of charge distribution[154], which can be properly captured by higher order multipole moments[155]. Therefore, tremendous efforts have been made to develop advanced representations of electrostatics ranging from fluctuating charges[156], Drude oscillators[157,158], up to fully polarizable force-fields such as AMOEBA[159,160].

Methods for binding free energy calculation can be classified according to depiction of either alchemical or physical pathways. The alchemical pathway uses alchemical, or non-physical intermediates to compute binding free energy, which is popular for its general applicability and efficiency. Physical pathways are preferable for large molecules and can give binding mechanism and kinetics[161,162]. While traditional methods such as Bennett's acceptance ratio (BAR)[163] have been successful, improvement in computational efficiency is desired for application to large systems and more

sophisticated potential energy representations. To this end, many enhanced sampling methods have been developed[164,165].

Host-guest systems are often used as a model for binding affinity prediction because of their modest size and high specificity among guests. By computing the free energy behaviour of relatively small molecules, inadequacies can be better determined and remediated for the rigorous and strenuous computation of binding free energies for large proteins. In the SAMPL3[166] and SAMPL4[167] host-guest binding competitions, the cucurbit[7]uril macrocycle was used as the host molecule. The cucurbit[n]uril macrocycle (CB[n]) is composed of *n* conjoined glycoluril subunits forming a cylindrical molecule approximately 9.1Å in height (for a thorough review of CB[n] chemistry, see [168]). As with many macrocycles, such as cyclodextrin, CB[n] has been explored as a molecular container for drug delivery[169-171]. The glycoluril subunits position a ring of carbonyl groups at the two faces of the cylinder, while the inner region of carbon-nitrogen chains remains hydrophobic. Hence, guests of hydrophobic cores with cationic end groups can bind with high affinity to the CB[n] host.

In this work, we report the investigation of host-guest binding thermodynamics between a CB[7] host and a set of 14 small molecules. The guests range from linear hydrocarbons to cycloalkanes, species of norbornanes and adamantane. We use two free energy calculation methods and several thermodynamic inquiries to interpret experimental affinities. In particular, we dissect the roles of entropy and enthalpy in binding for each guest. For anomalous enthalpy/entropy values, the separate entropy contributions of water and the host-guest systems are investigated. We determine that binding affinities of the host-guest systems are both enthalpy- and entropy-driven. We further discuss the application and convergence of the OSRW and BAR binding free

energy methods. Our results attest to the application of binding free energy simulation methods towards the understanding of experimental binding affinities.

3.2 METHODS

AMOEBA polarizable force field. An implementation of the polarizable AMOEBA force field with the molecular modeling software package, TINKER,[172] is used in this study. Typical force fields treat charges as static entities, usually represented as fixed-atomic charges.[152,173] However, the actual charge distributions of atoms change in response to the environment's electric field.[152,173-175] As a physics-grounded force field, AMOEBA depicts molecular polarizability and electrostatic potential terms by using mutual atomic dipole-dipole induction along with permanent atomic monopole, dipole, and quadrupole moments[159]. This results in a more accurate description of molecular energetics in protein-ligand binding. Although the polarizability of the CB[7] cavity is posited to be low[176], the authors anticipate that the large number of heavy atoms of the system as well as the differences in electronegativity of the guests make this host-guest system appropriate for employment of a polarizable force field.

Bennett acceptance ratio (BAR). BAR[163] is a method to calculate the free energy difference between different thermodynamic states. It has been shown to be more efficient than other classic methods such as free energy perturbation.[177] Typically, simulations are conducted at multiple intermediate states that connects the two end states, and free energy difference between neighbour states are calculated based on the energy difference.

Orthogonal space random walk (OSRW). OSRW is an enhanced sampling scheme for free energy calculation, which performs a random walk in two orthogonal dimensions.[178-181] One dimension is along the order parameter λ representing an

alchemical intermediate state that connects the two states of interest.[182,183] The other dimension is along the orthogonal generalized force ($F_\lambda = \partial U/\partial\lambda$), whose integral is the free energy (Eq. 1). Once a state is sampled, a Gaussian distributed bias is added to discourage the system from revisiting that state. The main advantage of OSRW over many other methods including BAR is that it accelerates the sampling of the orthogonal generalized force. A complete explanation of the method as well as the requisite adjustments needed to employ a polarizable force field can be found in Abella et al.[184]

$$\Delta G = \int_0^1 \langle \frac{\partial U}{\partial \lambda} \rangle_\lambda \, d\lambda \tag{1}$$

Absolute binding free energy. The absolute binding free energy ΔG_{bind} of the host-guest system was calculated by using the double-decoupling method, which refers to "disappearing" the guest in both solvent and a solvated host-guest complex.[185] As illustrated in Figure 3.S1, and articulated in Eq. 2, the binding free energy is the free energy difference between removing a guest from its water environment (ΔG_{hyd}) and decoupling a guest from its host–water environments (ΔG_{host}). At $\lambda = 0$ the guest is completely decoupled from its environment and can wander to other parts of the box, prolonging convergence. To solve this problem, a harmonic restraint[186] ($k = 15\ kcal/mol$) is added between the centers of mass of the host and the guest, and a correction term[185-187] is needed to recover the true free energy difference (Eq. 3).

$$\Delta G_{bind} = \Delta G_{host-guest} - \Delta G_{hydration} + G_{correction} \tag{2}$$

$$G_{correction} = RTln[C^o \left(\frac{2\pi RT}{k}\right)^{\frac{3}{2}}] \tag{3}$$

In Eq. 3, R is the gas constant, T is the temperature, C^o is the standard concentration and k is the force constant of the restraint. Finite size effects on charging free energies[188] were not corrected since they are expected to be similar for $\Delta G_{host-guest}$ and $\Delta G_{hydration}$ and will cancel out.

Computational Details. In this study, the absolute binding free energy values of 14 guests in the SAMPL4 CB[7]-guest system were calculated using the polarizable AMOEBA force field. Parameters for the molecules were derived by following the procedure previously described in Ren et al[160]. All molecular dynamics simulations were run using TINKER with a RESPA integrator[189] with a 2.0 femtosecond time step and Bussi thermostat[190]. The vdW calculations had a 12.0 Å cutoff while the electrostatics was calculated by particle mesh Ewald summation with a real-space cutoff of 7.0 Å. The Gaussian bias was deposited every 10 steps, with a height of 0.005 kcal/mol and widths of 4 kcal/mol for F_λ and 0.01 for λ. Additional simulations with a reduced height of 0.001 or 0.002 kcal/mol were also carried out for some guests. The production time of the OSRW is around 15-20 ns. All OSRW simulations were conducted on Texas Advanced Computing Center (TACC) Stampede as well as a local computer cluster. For the BAR simulations, first the electrostatics were gradually scaled off with vdW interactions kept at full strength, and then the vdW interactions were scaled off. The numbers of steps for these two stages were 11-12 and 10-13 respectively. The total simulation time for each step was 1 ns and coordinates were saved every 1 ps for analysis. The correction $G_{correction}$ was 6.245 kcal/mol and should be added to all binding free energy calculations for both BAR and OSRW. The uncertainties of the BAR results were estimated based on the distribution of uncorrelated samples, while the uncertainties of OSRW results were obtained by comparing the final values of independent simulations and are imprecise due to the small numbers of simulations. The binding enthalpy was obtained from the difference between the average energies in the binding and free states. This method has comparable accuracy with that of the van't Hoff method[191] and that of the BAR method[192].

3.3 RESULTS

Figure 3.1 and Table 3.1 both present binding free energy results from OSRW and BAR computations compared with experiment. In Table 3.1, structures and energies of the guest ligands studied here are presented[167]. The host for all ligands is CB[7] as stated previously. For each ligand in Table 3.1, the free energy values of experiment, OSRW, and BAR are presented explicitly. Reported in the SAMPL4 results, the absolute uncertainty of all experimental free energy values is ± 0.1 kcal/mol. The BAR results are those that were previously reported in the SAMPL4 contest[167].

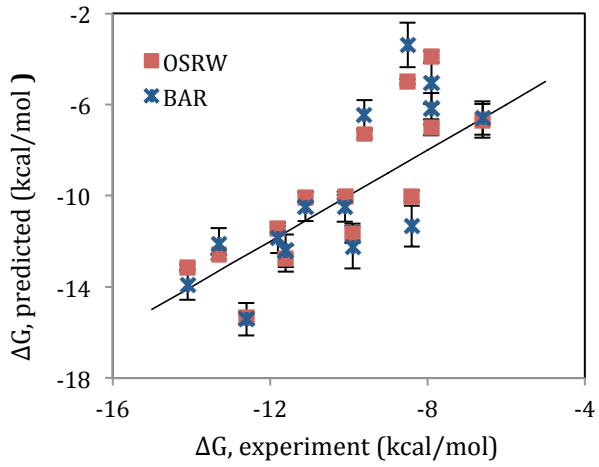


Figure 3.1: Predicted binding free energy as a function of experimental binding free energy (in kcal/mol). Line is y=x.

In Figure 3.1, both OSRW and BAR results are plotted against experimental free energy values taken from the SAMPL4 host-guest competition. The OSRW and BAR free energies establish good correlation with experiment, having R^2 correlation values of 0.69 (OSRW) and 0.62 (BAR). The OSRW and BAR results also agree with each other within statistical uncertainties for most of the systems. The discrepancies for the other

systems can be accounted for by the imprecise uncertainty estimate of the small numbers of OSRW simulations.

In Table 3.S1-3.S2 (Additional Figures and Tables, section 3.6), a decomposition of binding free energies is given. For ligand C5, positive binding free energies calculated from BAR led to the exploration of multiple ligand protonation states, denoted as C5 and C5b (see Figure 3.S7 for structure comparison). In five ligand cases (C1, C3, C5b, C9, and C10), the OSRW computation displayed large fluctuations in free energy. Since the fluctuation is proportional to the bias deposition rate, additional OSRW simulations were conducted with decreased Gaussian-height biases for each of these ligands. In theory, lowering the height of the Gaussian distribution will suppress fluctuations at the expense of slowing down dynamics. However, in this work, the OSRW computations with a lowered Gaussian height (LGH) bias converged at roughly the same simulation time as the original computations. Lastly, in Table 3.S1-3.S2, ligands C3 and C10 were duplicated in the OSRW computation due to poor convergence of the original simulations.

For Figure 3.1 and Table 3.1, the final OSRW ligand binding free energy value is taken to be the average over all of the OSRW computations for that ligand, with some values excluded (explained below). Multiple independent OSRW simulations were run for each ligand. As mentioned above, for two ligands, OSRW computations were repeated with the original parameters. The averaging of the free energies includes the LGH and repeated computations with the original pair of simulations. Exceptions to this average method are ligands C5 and ligands C10. The binding free energy value for ligand C5 was taken to be the average of the ligand C5b_LGH computation. The protonation state for ligand C5 reported by the BAR computations was similarly C5b. The 2.5 kcal/mol disagreement between original OSRW simulations for ligand C5b, as well as the

nice agreement between the C5b_LGH simulations (within 0.3 kcal/mol), supported our use of the C5b_LGH data. For ligand C10, all of the OSRW free energy values were used in the average except the -0.76 value as it was in disagreement with all of the other five values by 2.5 kcal/mol. We suspect that this low free energy value is an artefact of a slow-convergence binding energy computation.

Table 3.2 presents errors and correlation metrics between OSRW/BAR and experimental values. Despite the duplicate runs and Gaussian height decrease necessary for the OSRW computations, the Kendall τ coefficient for OSRW supports a strong agreement between OSRW and experiment. For further validation, in Figures 3.S2-3.S4, we have computed correlation metrics for all of the possible answer combinations from the host-guest and hydration free energy values in Table 3.S3. These three figures show the distribution of possible answer choices as well as our reported value and mean of possible answer choices. The OSRW computation times needed for all ligands range from 13.76 – 23.95 ns, further elucidated in the discussion section. The computational expense was likewise heavy, with several weeks required using the Texas Advanced Computing Center, as well as several weeks on a local cluster.

Table 3.1: Host-guest binding free energies. The OSRW column presents the average of results from the full length of simulations, while the OSRW (10 ns) column presents values at 10ns. The host molecule for all structures is cucurbit[7]uril. All free energies are given in kcal/mol. The experimental free energies hold an uncertainty of ± 0.1 kcal/mol.

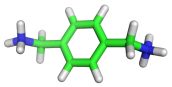
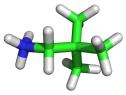
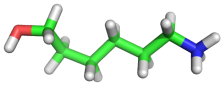
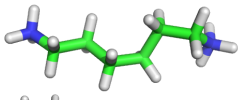
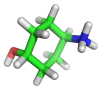
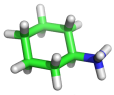
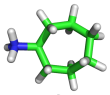
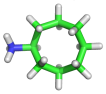
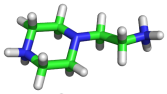
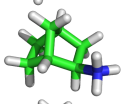
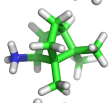
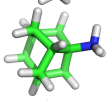
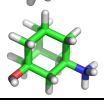
Guest	Guest Structure	ΔG_{bind}			
		Expt.	BAR	OSRW	OSRW (10 ns)
C1		-9.9	-12.27 ± 0.92	-11.64 ± 0.42	-10.43 ± 1.01
C2		-9.6	-6.46 ± 0.65	-7.29 ± 0.20	-5.50 ± 0.05
C3		-6.6	-6.59 ± 0.74	-6.71 ± 0.74	-5.17 ± 1.03
C4		-8.4	-11.34 ± 0.89	-10.02 ± 0.25	-8.31 ± 0.75
C5b		-8.5	-3.39 ± 0.97	-5.00 ± 0.11	-4.46 ± 0.26
C6		-7.9	-6.18 ± 0.69	-7.01 ± 0.35	-6.64 ± 0.07
C7		-10.1	-10.49 ± 0.66	-10.05 ± 0.09	-10.96 ± 0.78
C8		-11.8	-11.84 ± 0.68	-11.44 ± 0.09	-11.15 ± 0.76
C9		-12.6	-15.42 ± 0.71	-15.35 ± 0.29	-15.44 ± 0.65
C10		-7.9	-5.06 ± 0.91	-3.90 ± 0.26	-3.69 ± 0.68
C11		-11.1	-10.48 ± 0.64	-10.06 ± 0.25	-9.82 ± 0.34
C12		-13.3	-12.11 ± 0.70	-12.57 ± 0.03	-12.33 ± 0.16
C13		-14.1	-13.92 ± 0.65	-13.13 ± 0.13	-12.63 ± 0.52
C14		-11.6	-12.41 ± 0.72	-12.75 ± 0.59	-12.05 ± 0.58

Table 3.2: Model deviation from experiment. RMS energy difference, and AUE (Average Unsigned Error) are in kcal/mol.

Method	RMS Error	AUE	R^2	Kendall τ
OSRW	1.92	1.51	0.69	0.74
BAR	2.26	1.73	0.62	0.58
OSRW(10ns)	2.22	1.73	0.73	0.75

3.4 DISCUSSION

Analysis of computed binding affinities from the SAMPL challenge allows for elucidation of binding thermodynamics as well as examination of computational predictions. In the official SAMPL4 host-guest paper, free energy values from BAR simulations using the AMOEBA polarizable force field were noted for good performance[167]. Our OSRW-computed free energies correlate with experimental values slightly better than the BAR results. Note that both methods use the exact same parameter sets and simulations parameters. However, long computational times needed for convergence of OSRW free energy were observed. Upwards of 20 ns of computation time in binding was required for some ligands, while in our previous work, the hydration free energy was able to converge in less than 4 ns[184]. For comparison, the BAR computations were performed for 1ns for each vdW and electrostatic window. One possible reason that the OSRW method applied here may be slow to converge is due to the underlying metadynamics procedure. Recently, the Orthogonal Space Tempering[180] method has been proposed to address this problem.

We also investigated the OSRW results if the free energy computations were carried out for only 10 ns rather than continued to 15+ ns. In Tables 3.1-3.2 the 10ns OSRW binding free energies are presented in comparison to the experimental values. Surprisingly, the R^2 correlation value and the Kendall τ correlation coefficient are high, supporting strong correlation between OSRW and experiment after just 10ns simulations

(Table 3.2). Despite this strong correlation, the individual ligand errors and the RMSE between experiment and OSRW are slightly higher than the final results.

To gain insights into the molecular driving forces for binding, we have examined the enthalpy and entropy contributions of the binding free energy. Table 3.3 lists the calculated binding enthalpy and entropy for each guest ligand. Although the binding free energies for different ligands are close, ranging from -15 to -5 kcal/mol, the binding enthalpies are vastly different. This is a good demonstration of the enthalpy-entropy compensation in host-guest binding. Due to the relatively short simulation time (1 ns), the uncertainties are on the order of 10 kcal/mol. Nevertheless, it can be seen that some of the recognitions are driven by enthalpy while others by entropy. Ligands 9, 12, and 13 have both favourable binding enthalpy and entropy. Extreme examples are ligand C10 for entropy-driven binding, and ligands C7 and C8 for enthalpy-driven binding. However, there appears to be no simple relationship between the binding thermodynamics of the ligand and its charge or geometry. Comparing C5 with C5b and C3 with C4, we find that the binding enthalpy does not correlate with the net charge. Ligands C7, C8 and C9 have the same functional groups and their binding affinities increase with ring size, but their entropy values differ. Enthalpy values of ligands C7 and C8 clearly indicate a dominant contribution of enthalpy, while for ligand C9 the enthalpy value is competitive with entropy.

Further analyses were carried out to look into the binding mechanisms. To explain why guest ligands C7 and C8 are enthalpy-driven, we investigated the ligand hydrogen bonding formation in water and complexes. Table 3.4 lists hydrogen bond (H-bond) numbers for ligands C7, C8 and C10 between guest-water in solution and between guest-host/water in the complex. The data are averaged over 1000 frames in 1 ns. Compared to ligands C7 and C8, ligand C10 formed more H-bonds when free in water (5.7) and bound

to the complex (5.4). Furthermore, we analysed the portion of H-bonds formed between guest-host and guest-water. The three ligands formed similar numbers of H-bonds with the host while ligand C10 has twice the H-bonds formed with the surrounding water than other ligands. This may be attributed to the structural differences: ligand C10 has 3 polar amine groups with two of them exposed to the surroundings, attracting water and other polar groups. In contrast, ligands C7 and C8 have only one amine group each, leading to less intermolecular interaction. Noticeably, an increase of H-bonds in ligands C7 and C8 is found when moved from solution to the host-guest complex. On the other hand, the number of H-bonds formed by C10 decreases upon binding. The changes in H-bonds may explain why the binding of ligands C7 and C8 were found to be enthalpy-driven.

Table 3.3: Host-guest binding enthalpies and entropies (kcal/mol). *STD(ΔH)* is the uncertainty of enthalpy.

Molecule	*ΔH*	*STD(ΔH)*	*-TΔS*
C1	-14.91	13.87	2.64
C2	-17.39	13.54	10.93
C3	-18.58	14.79	12.00
C4	-6.62	14.02	-4.72
C5	3.03	15.24	-6.41
C5b	-5.08	13.96	1.53
C6	-12.48	14.60	6.30
C7	-26.99	13.47	16.56
C8	-28.56	14.30	16.72
C9	-3.69	13.41	-11.73
C10	45.20	13.71	-49.57
C11	1.33	13.37	-11.94
C12	-5.58	12.68	-6.67
C13	-7.53	14.19	-6.18
C14	4.28	12.63	-16.90

Table 3.4: Analysis of hydrogen bond numbers for guests C7, C8 and C10. The number of hydrogen bonds between guest-water in solution and between guest-host/water in host-guest complex are listed as $N_{solution}$ and $N_{complex}$ respectively. Further decompositions of hydrogen bond numbers between guest-host, and between guest-water in host-guest complex are given in $N_{complex}^{g-h}$ and $N_{complex}^{w-h}$ The presenting hydrogen bond numbers are averaged by 1000 frames over 1 ns.

Guest	$N_{solution}$	$N_{complex}$	$N_{complex}^{g-h}$	$N_{complex}^{g-w}$
C7	2.690	3.554	1.617	1.937
C8	2.792	3.389	1.201	2.188
C10	5.709	5.452	1.325	4.127

The rotation of guest ligands C7, C8, and C10 inside the CB[7] host was measured to explore the entropic aspects of these ligands. Three atoms from each ligand's aromatic ring were chosen to represent one plane, while three atoms from the host were chosen to produce a plane that bisects the host equally. The rotation of the guest plane with respect to the host plane was measured over the coordinates of 5ns trajectories. The potential of mean force (PMF) was also computed for the rotation angles. Similar to a study of an octa-acid host-guest system[144], the guest ligands here were determined to rotate almost freely with only small free energy barriers (~0.5 kcal/mol). Likewise, computation of the entropy using $S = -k_B * \Sigma pln(p)$ resulted in minute contributions ($S_{lig(complex)}^{rot} - S_{lig(free)}^{rot} \approx 0.05$ kcal/mol at T=300K), shown in Table S2.

The configurational entropies of host-guest complexes C7, C8, and C10 were computed using quasiharmonic analysis[193,194]. In the quasiharmonic analysis method, the mass weighted covariance matrix of atomic fluctuations is computed. Eigenvalues λ_i of this covariance matrix are then expounded to frequencies of collective motions, $\omega_i = (RT/\lambda_i)^{1/2}$. The estimated entropy S of the molecule is determined by Eq. 5 where R is the gas constant, $\hbar$ is Planck's constant, and T is temperature.

$$S = R \sum_{i=1}^{3N-6} \frac{\hbar\omega_i/RT}{e^{\hbar\omega_i/RT} - 1} - \ln\left(1 - e^{-\hbar\omega_i/RT}\right) \quad (5)$$

The quasiharmonic entropy was computed using AMBER14[148]. For each molecule, all heavy atoms (C,N,O) were included in the covariance matrix. **Table 5** shows the quasiharmonic entropy values for the host-guest systems C7, C8 and C10. These values include entropies of the host-guest complex S_{hg}, the guest only in complex $S_{g(complex)}$, the host only in complex $S_{h(complex)}$, the guest in solution $S_{g(solution)}$, the host in solution $S_{h(solution)}$, and the entropic contribution of binding $-T\Delta S_{conf}$ where $\Delta S_{conf} = S_{hg} - S_{h(solution)} - S_{g(solution)}$. The quasiharmonic approximation maintains limitations involving the use of Cartesian coordinates and the presence of multiple steep energy wells[195]. Further, the quasiharmonic approximation is known to present an upper-bound to entropy primarily due to correlations between modes[196,197]. However, several trends may be observed from the computed values. CB[7]-C10 complex has the highest entropic cost ($-T\Delta S_{conf}$) out of the three complexes computed.

Table 3.5: Configurational entropy computed from quasiharmonic analysis.[a] $S_h(solution)$ is 495.61 cal/mol/K.

Guest	S_{hg}	$S_{g(complex)}$	$S_{h(complex)}$	$S_{g(solution)}$	$-T\Delta S_{conf}$
C7	364.38	74.53	302.73	80.99	5.69
C8	379.41	91.88	289.87	96.34	3.12
C10	366.28	87.67	294.41	94.59	6.61

[a] Entropy values are given for S_{hg} the host-guest complex, $S_{g(complex)}$ the guest only in complex, $S_{h(complex)}$ the host only in complex, and $S_{g(solution)}$ the guest in solution. S_{hg}, $S_{g(complex)}$, $S_{h(complex)}$, and $S_{h(solution)}$ are computed from 5ns simulations while $S_{g(solution)}$ values are computed from 1ns simulations. All entropy values (except where marked) in cal/mol/K. $\Delta S_{conf} = S_{hg} - S_{h(solution)} - S_{g(solution)}$. $T\Delta S_{conf}$ computed at 300K, with units of kcal/mol.

Given that the enthalpy/entropy decomposition analysis suggested binding of guest C10 to be entropically favorable ($-T\Delta S_{tot} < 0$), the positive configurational entropy

change computed here (Table 3.5) indicates that the favourable binding entropy of ligand C10 is likely water driven. Binding of guests C7 and C8 resulted in approximately the same entropic cost. Although the values of S_{hg} and $S_{g(solution)}$ differ for guests C7 and C8, when combined, the values largely offset the differences. From analysis of guests C7 and C8, intramolecular atomic fluctuations of the aromatic carbon atoms inside the host are greater for C8 than for C7. This is consistent with intuition: the larger aromatic molecule of ligand 8 is slightly pressed inward by the host. This effect is evident in the $S_{h(complex)}$ values, where the host in guest C7 complex has roughly 4 kcal/mol greater entropy (*TS*) than the host in guest C8 complex, which is strained due to ligand size. Similar to the C10 complex, guests C7 and C8 complexes have a positive (unfavourable) entropy contribution, and additional unfavourable entropic interactions from water likely increase the binding entropy to the values in Table 3.3.

As noted above, there are discrepancies between OSRW and BAR results as well as between independent OSRW simulations for some ligands. To explain this, we observe that for an unbiased estimator, the uncertainty of a measured quantity is related to the sample distribution and the autocorrelation time as[177]

$$\sigma(\bar{A}) = \sigma(A)\sqrt{\frac{2\tau}{t}} \tag{6}$$

where τ is the integrated autocorrelation time and t is the total sampling time. $t/2\tau$ is also interpreted as the effective number of independent samples. Eq. 6 is valid for BAR. As for OSRW, since the underlying metadynamics does not converge asymptotically[198], Eq. 6 should provide a lower bound for its error. The sample distribution depends on the hybrid Hamiltonian, i.e. the decoupling scheme for the alchemical transition, which is different in the OSRW and BAR simulations. The correlation time varies with the simulation method. Generally, the correlation time in metadynamics should be shorter

than that of a classical molecular dynamics simulation on the same Hamiltonian. However, it is difficult to compare the correlation time between OSRW and BAR because OSRW has an additional degree of freedom λ. So here we focus on the effect of the decoupling scheme on the convergence. Figure 3.2 shows the standard deviation of F_λ for decoupling of guest C10 from its host-guest complex state in different decoupling schemes. When only the vdW interaction is decoupled (scaled down), the distribution F_λ is very narrow up to λ = 0.5. σ(F_λ) increases sharply and then falls to roughly 10 kcal/mol when λ goes from 0.5 to 0.6. When the electrostatics interaction is decoupled in the presence of vdW interaction, σ(F_λ) is nearly constant around 10 kcal/mol, which means that there is no dramatic change in phase space and that the evenly spaced λ points perform very well in distributing the simulation time. When vdW and electrostatics interactions are decoupled simultaneously, σ(F_λ) is significantly higher than when the two interactions are decoupled separately as λ approaches 0 and 1. In other words, decoupling the two interactions together enlarges the available phase space. As a result, more independent samples are needed for $\langle F_\lambda \rangle$ to converge at these two end states. In addition, we note that the autocorrelation time in the fixed λ OSRW simulations at λ = 0 is ~30 ps, much longer than in the BAR simulations when λ = 0. Based on Eq. 6, the uncertainty of the fixed λ OSRW simulations with a total simulation time of 20 ns was estimated to be ~1 kcal/mol. Although the dynamics are different from those of the OSRW simulations reported in Table 3.1, this result manifests that decoupling both interactions will create a rough energy landscape that makes sampling difficult. Therefore, the poor convergence of some of the OSRW simulations can be largely attributed to the decoupling scheme.

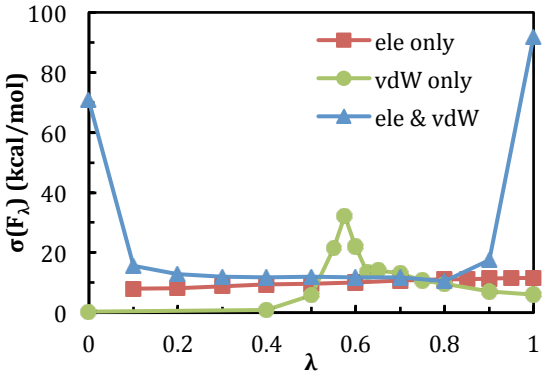


Figure 3.2: Standard deviation of F_λ as a function of λ for different coupling schemes. All analyses are based on the decoupling of guest C10 from its host-guest complex state. "vdW only" means that the vdW interaction is decoupled when there is no electrostatics. "ele only" means that the electrostatics is decoupled while vdW interaction is modelled at full strength. "ele & vdW" means that both electrostatics and vdW interactions are decoupled simultaneously as in the current OSRW implementation.

There is a positive correlation between the uncertainties of the OSRW simulations and the net charge of the system. Since the uncertainty estimate for OSRW results is limited in precision by the small number of simulations, here we use the differences between OSRW and BAR results to measure the uncertainties. Except for guest 3, all the OSRW results for systems with charge +1 agree well with those of BAR results, whereas large differences can be found for systems with charge +2 (See Table 3.6). This further supports our finding that decoupling vdW and electrostatics interactions together hinders the sampling. We expect that the problem will be less prominent for neutral systems.

Table 3.6: Correlation between uncertainties of binding free energies and net charge for each system. RMSE is the root mean square difference between OSRW results and the reference BAR results.

Guest	RMSE (kcal/mol)	Charge
C5b	1.35	2
C4	1.34	2
C10	1.28	2
C1	1.05	2
C5	0.98	1
C6	0.90	1
C2	0.85	1
C13	0.80	1
C3	0.80	1
C14	0.68	1
C11	0.49	1
C12	0.46	1
C7	0.45	1
C8	0.41	1
C9	0.30	1

3.5 CONCLUSION

In this work, binding free energies of the SAMPL4 host-guest system CB[7] with 14 guest molecules were computed with both BAR and OSRW methods and AMOEBA polarizable force field. Overall the AMOEBA binding free energy values computed using both BAR and OSRW are in good agreement with experimental results. The binding thermodynamics of this series of host-guest systems varies from ligand to ligand. Some are driven by enthalpy changes while others by entropy gains. We further examined guest ligands C7, C8 and C10, which display high enthalpy or entropy changes upon binding. The enthalpy-entropy decomposition suggests that the binding of guest C10 is entropy driven, while binding of guests C7 and C8 have large enthalpic contributions. Hydrogen bonding analysis showed that guest C10 formed several hydrogen bonding interactions with both water and host CB[7], largely due to the three hydrophilic amine groups.

Guests C7 and C8 gain additional H-bonds upon binding while C10 loses H-bonds upon binding, consistent with the enthalpy-entropy decomposition results. Configurational entropy was computed for guests C7, C8, C10 and their complexes with the host using quasiharmonic analysis. The configurational binding entropy was determined to be relatively small for all guests, hinting at the substantial role of water molecules. Through analysis of intramolecular atomic fluctuations of guests C7 and C8, cyclic carbon atoms inside the host were found to fluctuate more for guest C8 than C7, intuitively a result of the larger ring of C8. Unlike ligand-protein binding, the guest molecules were observed to freely rotate inside the host ring. Convergence of the BAR and OSRW free energy calculation methods were compared. The current OSRW implementation encounters convergence problems at the low end of vdW and electrostatics decoupling. Possible improvements can be achieved by separating the vdW and electrostatic decoupling, well-tempered metadynamics[198] and employing metadynamic alternatives[180]. Nonetheless, here, both BAR and OSRW methods are found to be adequate to determine the binding affinities for the model host-guest systems.

3.6 ADDITIONAL FIGURES AND TABLES

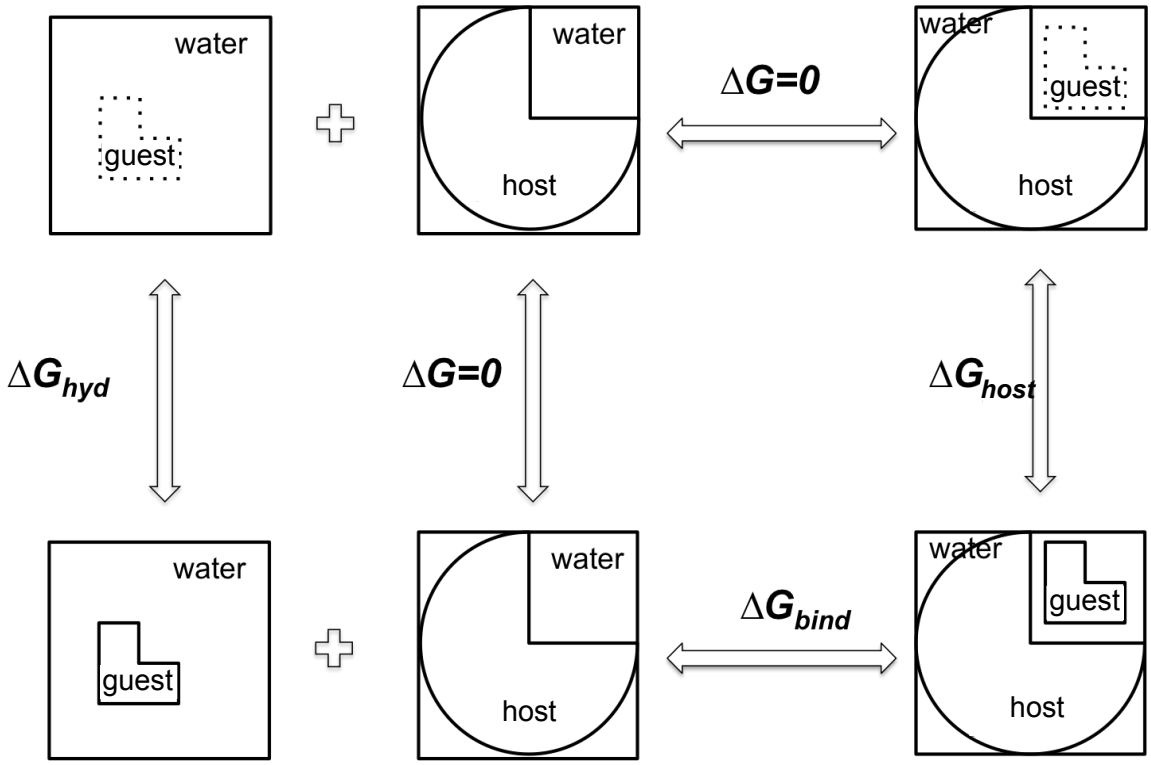


Figure 3.S1: Thermodynamic cycle for calculating the binding free energy of the host-guest system. The binding free energy (ΔG_{bind}) is defined as the difference between the decoupling free energies from both solvent and solvated protein complex. ΔG_{host} indicates that the ligand is decoupled from its protein environment, and ΔG_{hyd} indicates that the ligand is removed from a water environment.

Table 3.S1: Free energy composition of host-guest systems. All guests bind to the cucurbit[7]uril host. All free energies are in kcal/mol.

	Host-guest Free Energy ($\Delta G_{host-guest}$)			Hydration Free Energy ($\Delta G_{hydration}$)		
Guest	OSRW 1	OSRW 2	BAR	OSRW 1	OSRW 2	BAR
C1	-189.58	-188.11	-189.19	-170.42	-170.59	-170.67
C1_LGH[a,b]	-187.87	-188.20	–	–	–	–
C2	-70.21	-70.57	-70.20	-56.87	-56.84	-57.50
C3	-65.43	-64.33	-71.29	-52.42	-52.71	-58.45
C3_2[c]	-63.87	-65.42	–	-52.23	-52.93	–
C3_LGH[a]	-67.78	-64.75	–	-51.28	-52.28	–
C4	-180.21	-180.12	-181.27	-164.19	-163.61	-163.68
C5[d]	-68.64	-67.61	-68.01	-57.64	-58.10	-58.39
C5b[d]	-208.76	-207.18	-213.19	-199.57	-200.50	-203.55
C5b_LGH[a,d]	-211.84	-210.32	–	-200.48	-199.18	–
C6	-77.00	-76.25	-76.04	-63.39	-63.34	-63.62
C7	-74.25	-74.54	-74.69	-58.04	-58.15	-57.96
C8	-73.59	-73.48	-73.77	-55.81	-55.89	-55.68
C9	-77.98	-77.60	-77.98	-55.86	-56.78	-56.32
C9_LGH[a,e]	–	–	–	-56.07	-56.05	–
C10	-182.05	-186.04	-186.17	-175.04	-175.11	-174.86
C10_2[c]	-184.24	-183.61	–	-173.67	-174.09	–
C10_LGH[a]	-183.17	-182.70	–	-173.48	-172.69	–
C11	-71.68	-71.19	-71.69	-55.12	-55.14	-54.97
C12	-69.23	-69.89	-68.91	-50.39	-51.10	-50.56
C13	-71.08	-70.59	-71.56	-51.58	-51.34	-51.39
C14	-79.05	-77.97	-78.05	-59.47	-59.46	-59.39

[a] For all LGH ligand simulations, the height of the Gaussian bias (preventing OSRW from resampling that point) was lowered in order to promote more stable sampling. Not applicable to BAR simulations. [b] For C1_LGH, the host-guest free energies were coupled with C1 hydration free energies to compute the binding free energy. [c] C3_2 and C10_2 were independent OSRW simulations with the same system and parameters as C3 and C10. [d] C5 and C5b, are different protonation states of the C5 ligand. [e] For C9_LGH, the hydration free energies were coupled with C9 host-guest free energies to compute the binding free energy. [f] All experiment values hold absolute uncertainties of 0.1 kcal/mol. [g] These values were not included in the average for the reported results.

Table 3.S2: Binding free energy of host-guest systems. All guests bind to the cucurbit[7]uril host. All free energies are in kcal/mol. For BAR and OSRW, $\Delta G_{bind} = \Delta G_{host-guest} - \Delta G_{hydration} + G_{correction}$, where $G_{correction} =$ 6.245 kcal/mol.

	Binding Free Energy (ΔG_{bind})			
Guest	OSRW 1	OSRW 2	BAR	Experiment[f]
C1	-12.91	-11.18	-12.27	-9.90
C1_LGH[a,b]	-11.21	-11.26	–	
C2	-7.09	-7.48	-6.46	-9.60
C3	-6.76	-5.38	-6.59	-6.60
C3_2[c]	-5.39	-6.25	–	
C3_LGH[a]	-10.26	-6.22	–	
C4	-9.77	-10.26	-11.34	-8.40
C5[d]	-4.75[g]	-3.26[g]	-3.37	-8.50
C5b[d]	-2.94[g]	-0.43[g]	-3.39	
C5b_LGH[a,d]	-5.11	-4.89	–	
C6	-7.36	-6.66	-6.18	-7.90
C7	-9.96	-10.14	-10.49	-10.10
C8	-11.53	-11.35	-11.84	-11.80
C9	-15.87	-14.57	-15.42	-12.60
C9_LGH[a,e]	-15.66	-15.31	–	
C10	-0.76[g]	-4.68	-5.06	-7.90
C10_2[c]	-4.32	-3.27	–	
C10_LGH[a]	-3.45	-3.76	–	
C11	-10.31	-9.81	-10.48	-11.10
C12	-12.59	-12.54	-12.11	-13.30
C13	-13.25	-13.00	-13.92	-14.10
C14	-13.33	-12.16	-12.41	-11.60

[f] All experiment values hold absolute uncertainties of 0.1 kcal/mol. [g] These values were not included in the average for the reported results.

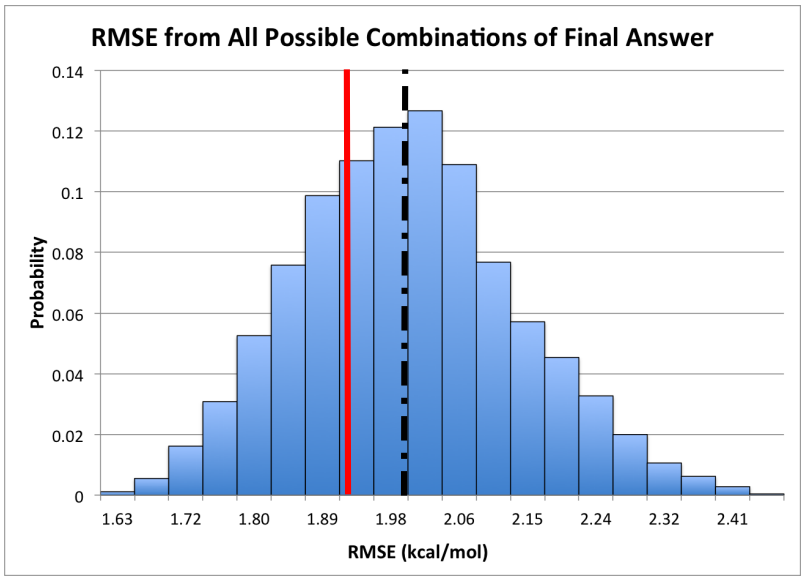


Figure 3.S2: Probability distribution of RMSE between calculated and experimental results from all possible OSRW answer combinations across all ligands. Solid red line represents the reported value in Table 2 of the main text, while the black dot-dashed line represents the mean value.

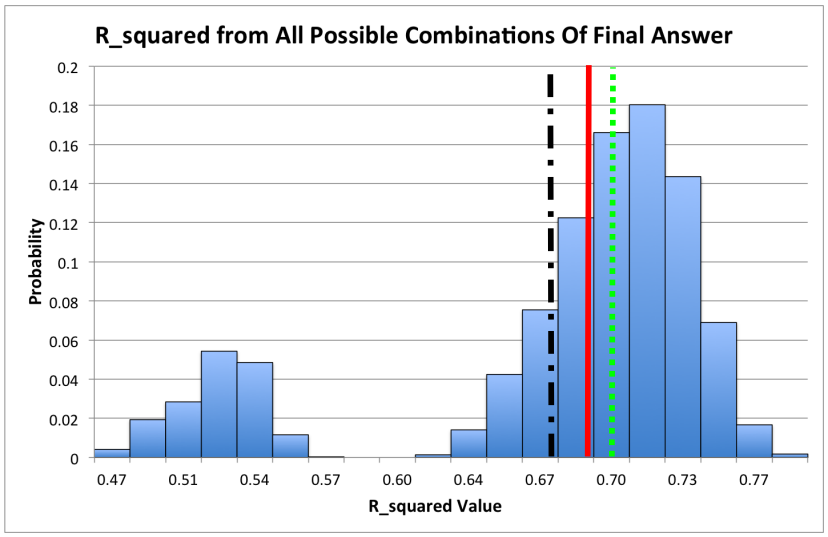


Figure 3.S3: Probability distribution of R^2 correlation coefficient from all possible OSRW answer combinations. Solid red line shows reported R^2 value. The black dot-dashed line represents the location of the mean while the green dotted line represents the median.

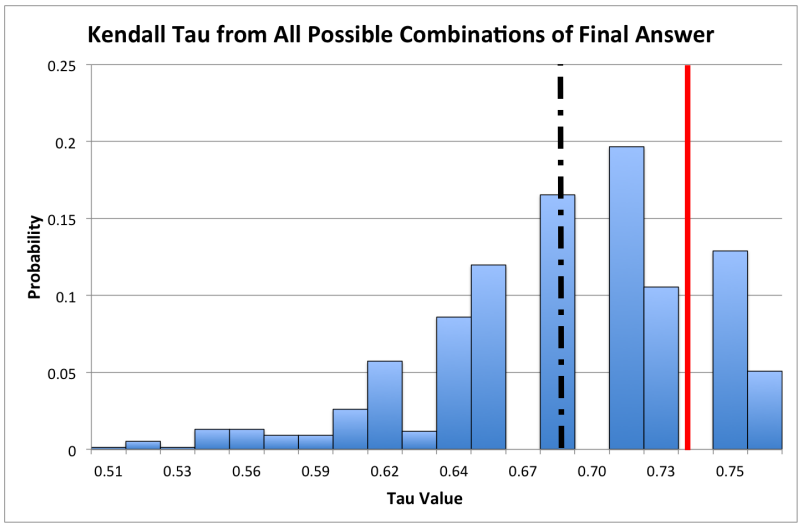


Figure 3.S4: Probability distribution of Kendall τ correlation coefficient from all possible OSRW answer combinations. Solid red line shows the approximate location of the reported τ value while the black dot-dashed line represents the mean.

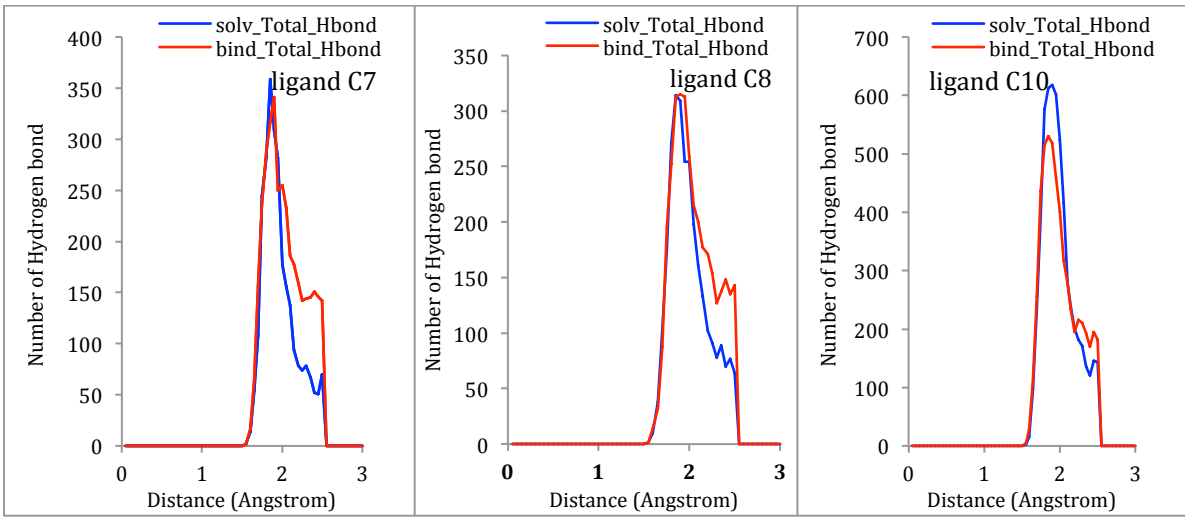


Figure 3.S5: Plots of the total number of hydrogen bonds for ligand C7, C8 and C10 between guest and water in solution (solv_Total_Hbond) and between guest and host/water in host-guest complex (bind_Total_Hbond).

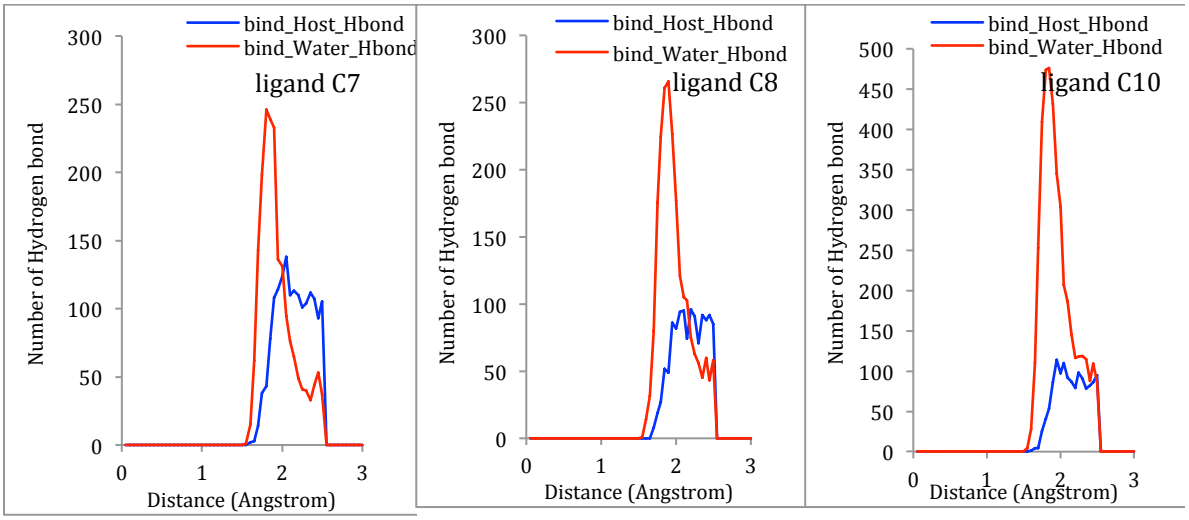
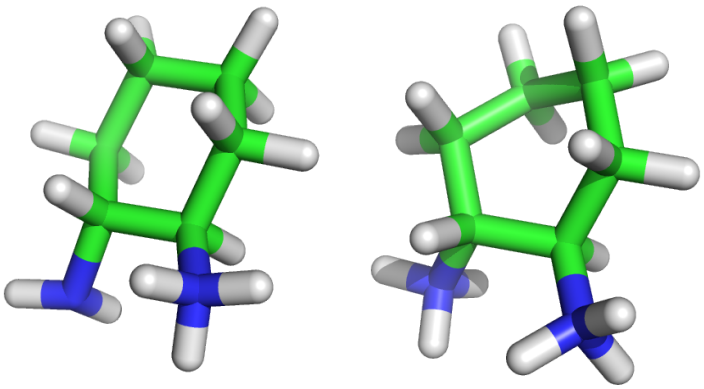


Figure 3.S6: Plots of hydrogen bond numbers between guest and host (bind_Host_Hbond), and between guest and water (bind_Water_Hbond) in host-guest complex.

Table 3.S3: Rotational entropy computed from PMF curves.

Guest ligand	$-T\Delta S$ (kcal/mol)[a]
C7	0.042
C8	0.053
C10	0.078

[a] $\Delta S = -k_B \Sigma p \ln(p) - S_{reference}$, where p is the probability taken from the PMF distribution and k_B is Boltzmann's constant. $S_{reference}$ is the entropy of the unbound state, taken as a uniform PMF distribution over the rotation angle bins, $S_{reference} = -k_B \sum \frac{1}{n} \ln\left(\frac{1}{n}\right)$, where n is the number of bins. T = 300K.



C5 C5b

Figure 3.S7: Structures of different protonation states of guest C5: C5 and C5b.

Chapter 4: Increased concentration of CdSe quantum dots increases specificity for PRM binding site of the SH3 domain

4.1 INTRODUCTION

Synthetic nanoparticles are employed in a multitude of applications. With increasing use, it is important to understand nanoparticle interactions with biological constituents. Carbon nanomaterials, such as graphene and bare carbon nanotubes (CNTs) have been determined to be highly destructive to cellular membranes[199,200] as well as cause substantial protein adsorption[201-206]. Toxicity is lessened in fullerenes and their derivatives[207,208], yet still present depending on surface modifications. In the special instance of gadolinium metallofullerenols, toxicity contributes to beneficial inhibition of tumor metastasis[209,210] as well as aiding uptake of therapeutics[211]. Hence, rigorous study of nanoparticle interaction with biological domains is needed for comprehensive application.

Cadmium selenide quantum dots (CdSe QDs) are photoluminescent nanoparticles employed in solar panels[212,213], consumer electronics, and biomedical imaging[214-218]. QDs are on the same size order of their Bohr radius allowing for generation of an exciton. QD emission can be tuned based on size, as larger QDs will have smaller band gap energies and thus longer emission wavelengths. QDs are used regularly in biomedical applications because of their broad UV absorption spectrum, narrow emission, and long-time photostability. A drawback of CdSe QDs is toxicity, which occurs both *in vitro* and *in vivo*[219,220]. QD toxicity is suspected to occur through three mechanisms: cadmium leaching, generation of reactive oxygen species, and interference with biomolecules[217]. Interference with biomolecules is suspected as cadmium leaching and reactive oxygen species cannot entirely explain observed toxicity[217,221].

Here, we investigate CdSe QD toxicity on the Src homology 3 (SH3) protein domain using all-atom Molecular Dynamics (MD) simulations. SH3 is a protein binding domain with over 250 instances in the human genome[222] and is essential to cell signaling and regulation[222-226]. We explore several different concentrations in order to understand the dose dependent interaction. Our results support conclusions drawn from experiment: that QDs exhibit a dose dependent and surface coating dependent toxicity.

4.2 METHODS

The SH3 protein structure used in this work was taken from a crystallized C-Crk, N-terminal SH3 domain in complex with the proline-rich SOS peptide(sequence: PPPVPPR), PDB ID: 1CKB[227]. The QDs used consisted of $(CdSe)_{13}$ nanoparticles, with ten trioctylphosphine oxide (TOPO) covalently bound to surface exposed Cd ions. Four initial system setups were used: a "binary" system with 1 QD and SH3; a "ternary" system with 1QD, PRM and SH3; a system with 2 aggregated QD (e.g. QD dimer) and SH3, and a system with 4 aggregated QD (e.g. QD tetramer) and SH3. All macromolecules (SH3,QD,PRM, excluding QD dimer and tetramer) were initially separated by 15Å. QD dimers and tetramers were simulated due to QD's tendency to aggregate in solution; dimers and tetramers were built by simulating separated quantum dots in solution, with aggregation occurring within 200ns. For each configuration setup, five different systems were generated by rotating the SH3 protein 72° about a vertical axis; additionally, 2 simulation runs were added for the QD tetramer configuration. In total, 32 systems were simulated. Each system was immersed in a roughly 87x87x87Å box of TIP3P water molecules. NaCl ions were added to all systems to achieve 0.1M salt concentration. The CHARMM36 force field[228] was used for protein parameters, while QD parameters were found from collaborators with TOPO parameters from

CHARMM22 lipid force field[229,230]. Van der Waals interactions were cutoff at 12Å while long range electrostatics was treated with the Particle Mesh Ewald[231] method. Systems were energy minimized for 15,000 steps with fixed protein atoms, followed by 5,000 steps of energy minimization with all atoms allowed to move. Following minimization, systems were equilibrated for 500,000 steps with a 0.5 fs timestep. Production Molecular Dynamics (MD) simulations were run for 200ns with a 2fs timestep. All simulations were run in the NPT ensemble, with a pressure of 1atm and temperature of 310K. In total, over 6.5 μs of all-atom MD was simulated. Simulations were run using NAMD2[232] on both an IBM Blue Gene Q[233] machine as well as an IBM Power 8 cluster.

Contact ratio is computed for all four systems studied, with a contact distance of 4.5Å. Contact ratio is defined as the number of frames the QD contacted the SH3 domain, divided by the total number of frames. For the aggregated dimer and tetramer systems, contact was calculated between the QD group and the SH3, not individually. Binding free energy surfaces were computed similarly to ref[234] by creating a 2-D probability histogram $P(S,D)$ of surface contact area *(S)* and key-residue distance *(D)*. The PMF was computed from $W(S,D) = -RT\ln(P(S,D))$ assuming a uniform reference distribution. The key binding site residues are F 141, W 169, P 183, and Y 186.

4.3 Results and Discussion

QD interactions with SH3 were explored with four different systems. A monomer system (M) with the SH3 domain and one QD was studied to understand general binding preferences for an isolated QD. A ternary system (Mt) with the SH3 domain, one QD, and the native PRM ligand (PPPVPPR) was investigated to capture competitive binding between the QD and PRM for the SH3 domain. To explore the effect of concentration on

SH3-QD interactions, a dimer system (D) with the SH3 domain and 2 QDs as well as a tetramer system (T) with the SH3 domain and 4 QDs were studied. For all systems, the CdSe QDs were coated in ten trioctylphosphine oxide (TOPO) chains. QDs with TOPO chains aggregate in solution and enter the cell through endocytosis as a conglomerated unit. Reflecting this, the dimer and tetramer systems started out and remained aggregated for this study. Five different configurations for each system were immersed in a TIP3P water box and explored using 200ns Molecular Dynamics (MD) simulations with the NAMD2 software package. The CHARMM36 protein force field was used for protein parameters, while QD-TOPO chain parameters used CHARMM22 lipid force field as well as solved parameters. Simulation details are discussed in the Methods section.

Figure 4.1a presents our most important finding: quantum dots have increasing preference for the SH3 PRM binding site with increasing QD concentration, even surpassing the contact affinity of the native PRM ligand. In the monomer system, the QD does not make any contacts with the PRM binding site. However, in the dimer and tetramer systems, the QD has a high affinity for the binding site. This agrees with experiment, where QD toxicity is found to be dose dependent. Our results support that at low concentrations, there is little contact with the active site and presumed little effect on function. At high concentrations, there is high affinity for the binding site and presumed large effects on function.

Shown in Figure 4.1b, the SH3 domain studied is a 56 residue β-barrel domain. The key PRM binding residues of SH3 are shown in green in Figure 4.1b and are SH3 residues F 141, W 169, P 183, and Y 186. The binding site is a narrow hydrophobic region, with a negatively charged "specificity pocket" of aspartic and glutamic acid interacting with the positively charged arginine of the PRM. In all simulations, the SH3 domain remained relatively stable with low RMSD (most < 2Å, max 4.2Å) and RMSF (<

3Å) as shown in the SI Figures 4.S1 and 4.S2. This is in contrast to carbon allotropes including carbon nanotubes and graphene, which have been shown to severely disrupt protein structure. One difference between the carbon allotropes mentioned and the QDs studied is the lack of cyclic rings with disperse electrons in the QDs.

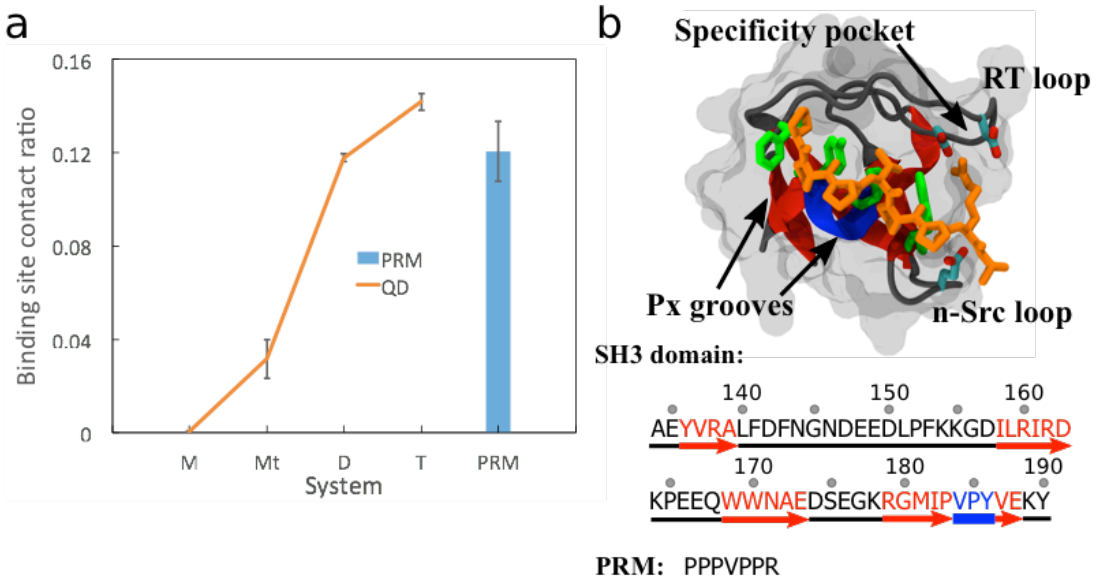


Figure 4.1: Main result. (a) Average QD-key residue (binding site) contact ratio over studied systems with PRM-key residue contact ratio for comparison. (b)Native SH3-PRM structure and sequence. PRM is shown in orange with the key PRM binding residues (SH3 141,169,183,186) shown in green. For the sequence, red arrows are beta strands, blue region is a 3/10 helix, and the black regions are loops. The RT loop spans residues 140-157 while the n-Src loop spans residues 164-168.

Figure 4.2 illustrates the initial configurations for each of the systems studied: monomer M, ternary Mt, dimer D, and tetramer T. Unlike fullerenes or single-walled CNTs which are smaller, the QDs studied here are of similar dimensions to the SH3 domain, especially with the TOPO chain surface coating. In the tetramer T system, the aggregated QDs form a tetrahedral structure to reduce solvent exposed surface area. As

mentioned in the methods, 5 different configurations were simulated for each system by rotating the SH3 domain equidistant about its vertical axis.

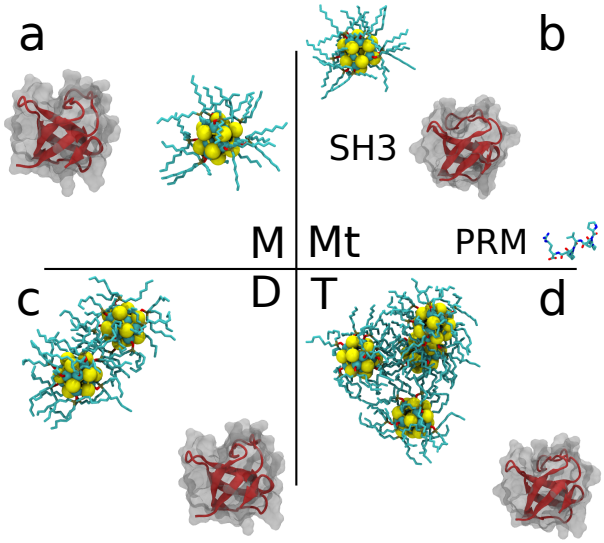


Figure 4.2: Initial system configurations. (a) Monomer system M, (b) ternary system Mt, (c) dimer system D, (d) tetramer system T.

Figure 4.3 shows the QD-SH3 residue contact ratio for all systems, with binding site residues indicated by black arrows. The monomer system has a high affinity to contact a distal site, residues 136 TYR, 158 ILE, and 175 SER. The ternary system favors the same site as the monomer system, yet with distraction by the PRM, the contact ratio is reduced. The dimer system shows modest affinity for the active site, but also maintains distal site contacts. The tetramer system shows the strongest specificity for the active site, with little contacts for the distal site, but favorable contacts with the binding site residues.

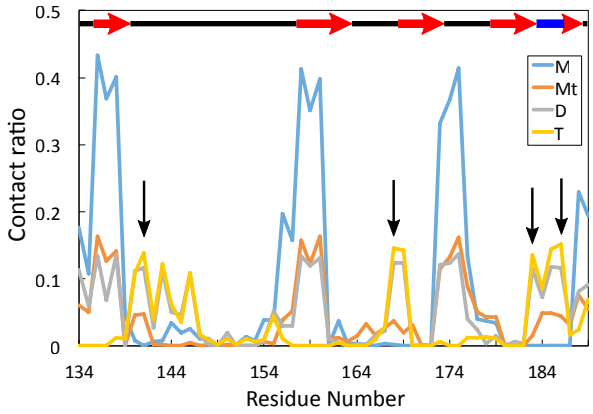


Figure 4.3: Contact ratio over all systems. Secondary structure is shown at the top, where red arrows indicate beta strands, blue bold line indicates 3/10 helix, and black lines indicate loop regions. Black arrows indicate binding site residues. Note that the monomer M and ternary Mt systems have little contact with the binding site but favorable contacts with distal site, while the dimer D and tetramer T systems have favorable affinity with the binding site.

Figure 4.4 presents the binding free energy surfaces and associated characteristic binding structures for the monomer, dimer, and tetramer systems. For the monomer system, no binding site affinity is observed. Instead, the QD shows high affinity for a site distal to the PRM binding site (blue well shown in Figure 4.4a). The distal site, shown in Figure 4.4b, is comprised of 136 TYR, 158 ILE, 175 SER. In general, the alkyl chains contact SH3 and have a high preference for hydrophobic residues. In one of the five configurations for this system, a QD is located 15 Å directly across from the binding site, yet the QD still bypasses the active site for contacts with the distal site. In the dimer case, the QD's can bind to both the distal site and the active site, yet there is a stronger affinity for the PRM binding site. Notably, the aggregated QD dimer can bind to SH3 as both a monomer and a dimer. The tetramer system only binds to the PRM binding site. The QD tetramer can bind to SH3 as a monomer, dimer, and trimer, but the tetrahedral structure

prevents all four QDs contacting the SH3 at the same time. For comparison, the PRM-SH3 binding free energy surface is shown in SI Figure 4.S3.

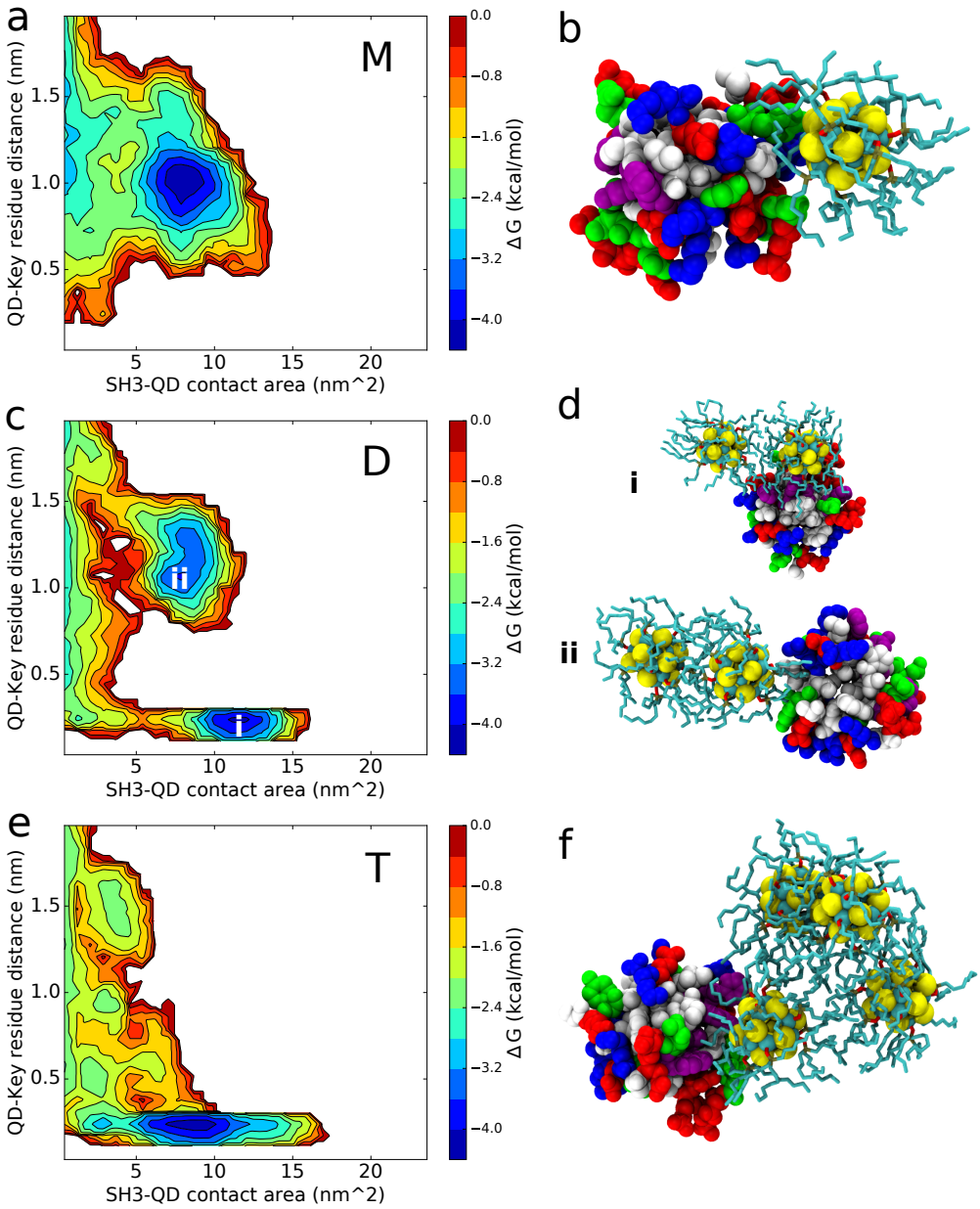


Figure 4.4: Binding surfaces and structures. (a,c,e) Binding free energy surfaces of (a) monomer system, M (c) dimer system, D and (e) tetramer system, T. (b,d,f) Characteristic structures of binding site well (blue area) for (b) monomer system, (d) dimer system, and (f) tetramer system.

Figure 4.5 depicts the binding competition between the QD and PRM in the ternary system. Shown in Figure 4.5a, the PRM clearly prefers contacting the binding site while the QD favors the distal site residues similar to the monomer system. The slight contact affinity of the QD for the binding site residues is due to the PRM "leading" the QD to the active site. The PRM has a strong preference to interact with the QD over the SH3 domain as observed in Figure 4.5b. From observing trajectories, the QD is able to steal the PRM from the SH3 binding site. The binding surface free energy of Figure 4.5c supports that the monomer QD favors binding with the distal site, and active site binding is not preferred but occurs due to the PRM interaction. As observed in Figure 4.5d, the QD is able to bind both the PRM and the SH3 domain simultaneously.

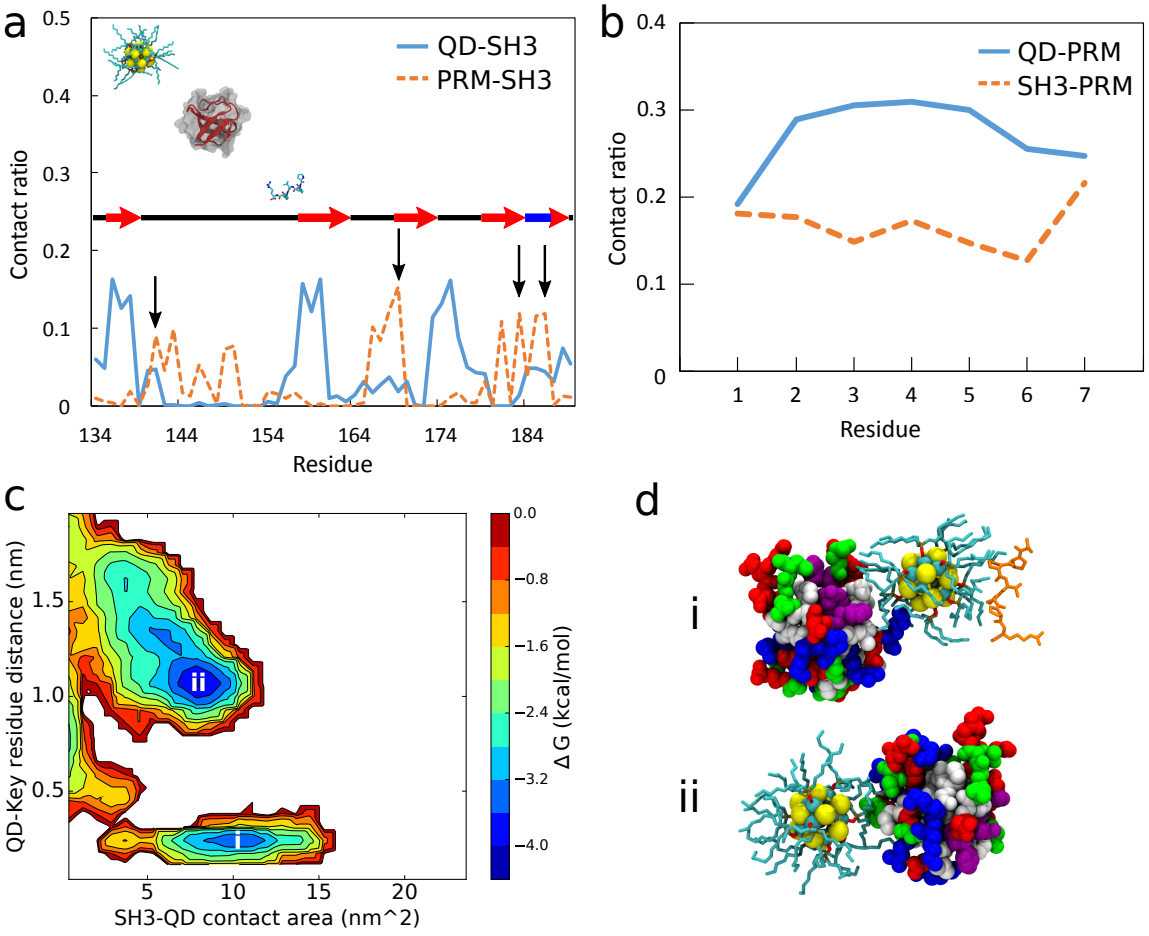


Figure 4.5: Ternary system. (a) QD-SH3 (blue, solid line) and PRM-SH3 (orange, dashed line) contact ratio over all frames. (a,insert) Ternary system initial configuration. (b) QD-PRM (blue, solid line) and SH3-PRM (orange, dashed line) contact ratio over all frames. (c) QD-SH3 Binding free energy surface as a function of QD-key residue distance and contact area. (d) Characteristic structures of main QD binding modes.

As all systems show interactions with SH3, it is important to consider the underlying QD structure that allows binding. The $(CdSe)_{13}$ core presented in SI Figure 4.S4 is composed of 13 Cd and 13 Se atoms. The Cd atoms retain a partial positive charge (Average=0.53) while the Se atoms retain a partial negative charge (Average=-

0.57). The net charge of the QD is -0.53. Figure 4.6 shows the sum of contact ratio between the QD core and protein for each system. For the dimer and tetramer systems, the QD cores were treated as a combined unit. The PRM has the highest preference to contact the QD core. As shown in Figure 4.6 inset, the binding mode of the PRM is for positively charged Arginine to contact the negatively charged Se atoms of the QD while the rest of the PRM interacts favorably with the TOPO chains. For the SH3 domain, contact with the QD core is opposite to that of the binding site contact trend, with the monomer system having the highest contact and the dimer and tetramer systems having decreasing contacts. As the QDs aggregate, the core becomes sequestered while the TOPO chains are pushed outward. It follows that as the TOPO chains are pushed outward, there is increasing preference to contact the PRM active site of SH3. This result provides evidence for surface coating dependent toxicity. The TOPO chains used here (and remain popular in experiments) are profoundly hydrophobic enabling beneficial interaction with the hydrophobic binding site. A hydrophilic surface coating ligand, such as polyethylene glycol (PEG), will decrease aggregation and will affect the protein binding affinity of the QD. This agrees well with experiment that shows decreased toxicity when the QDs are coated in PEG[235].

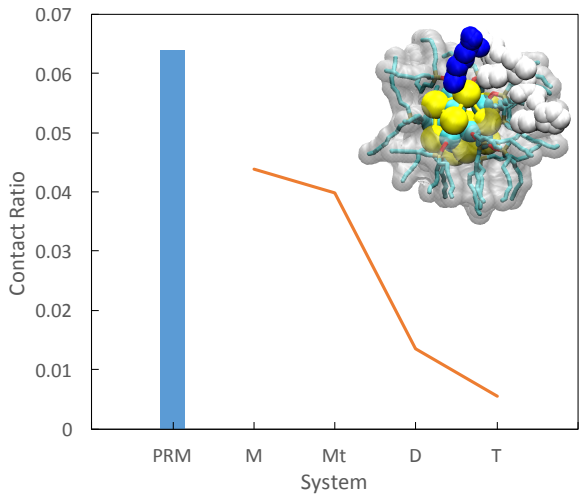


Figure 4.6: CdSe quantum dot interactions. Sum of the protein-contact ratio with the CdSe core. PRM interacts most readily with the CdSe core. For the SH3 systems, as the concentration and order of the quantum dots increases, the CdSe cores become sequestered with increasing pressure for TOPO exposure. (inset) Characteristic structure of PRM binding mode. The positively charged Arginine of the PRM interacts with the partially negatively charged selenium atoms while the hydrophobic residues prefer the TOPO chains.

4.4 Conclusion

At low concentrations, the QD specificity for the SH3 binding site is low, instead preferring a distal site. This corresponds to the SH3 domain retaining function/activity with PRM binding. With increasing concentration, there is increasing preference for the QD to contact the PRM binding site of the SH3 domain. With increasing QD concentration, the coating TOPO molecules are pushed outward to the surface, where they have preferable interactions with the hydrophobic PRM binding site. The large contact preference for the binding site with high QD concentrations indicates that the PRM will not successfully bind to the SH3 domain and the protein will lose function. This loss of function indicates that at high concentrations, the studied QDs are toxic. However, in all of our systems, the SH3 domain retains its structure with most staying

below 2 Å RMSD of protein backbone atoms to the crystal structure, and a max of 4.2 Å RMSD. Ultimately, we show that QD interaction with protein is heavily dependent on concentration and the choice of surface coating with varying toxicity.

4.5 Additional Figures

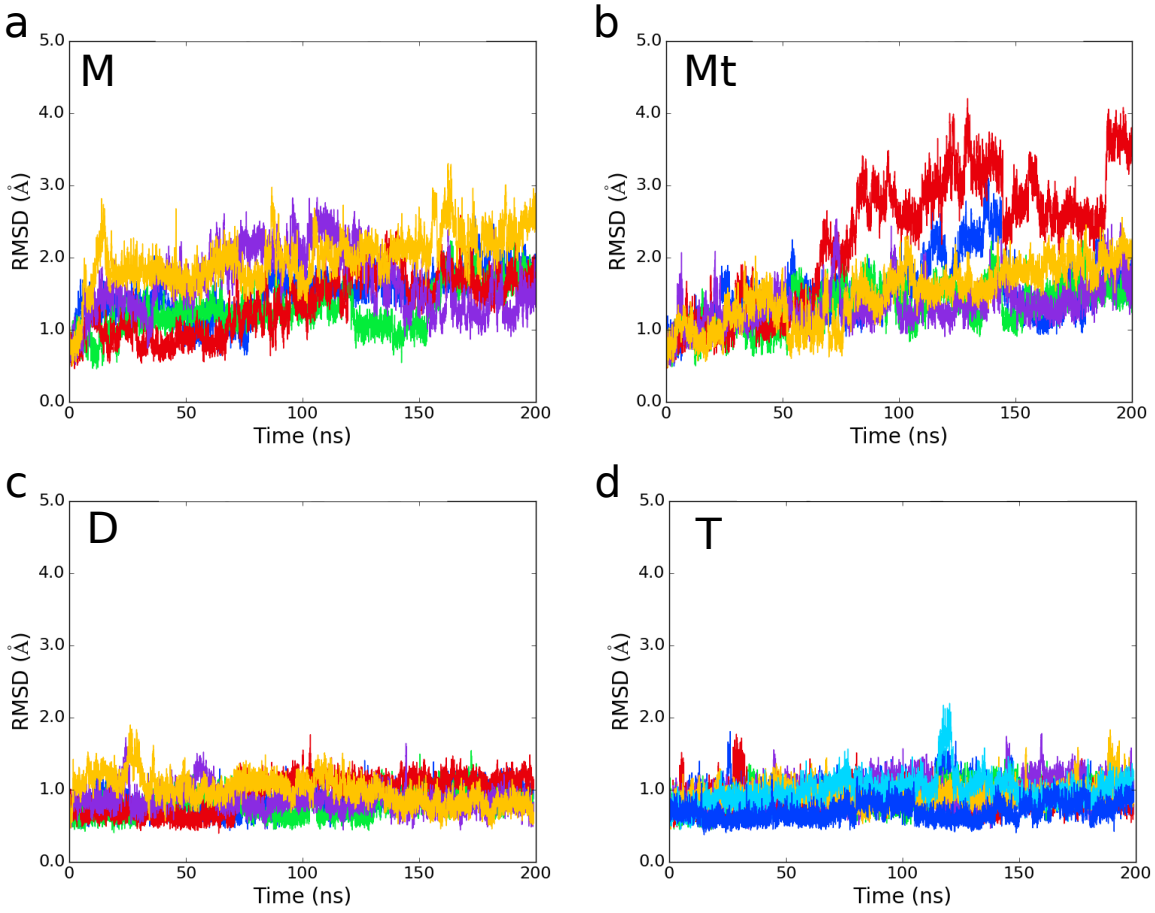


Figure 4.S1: RMSD of SH3 domain in the (a) monomer M, (b) ternary Mt, (c) dimer D, and (d) tetramer T systems.

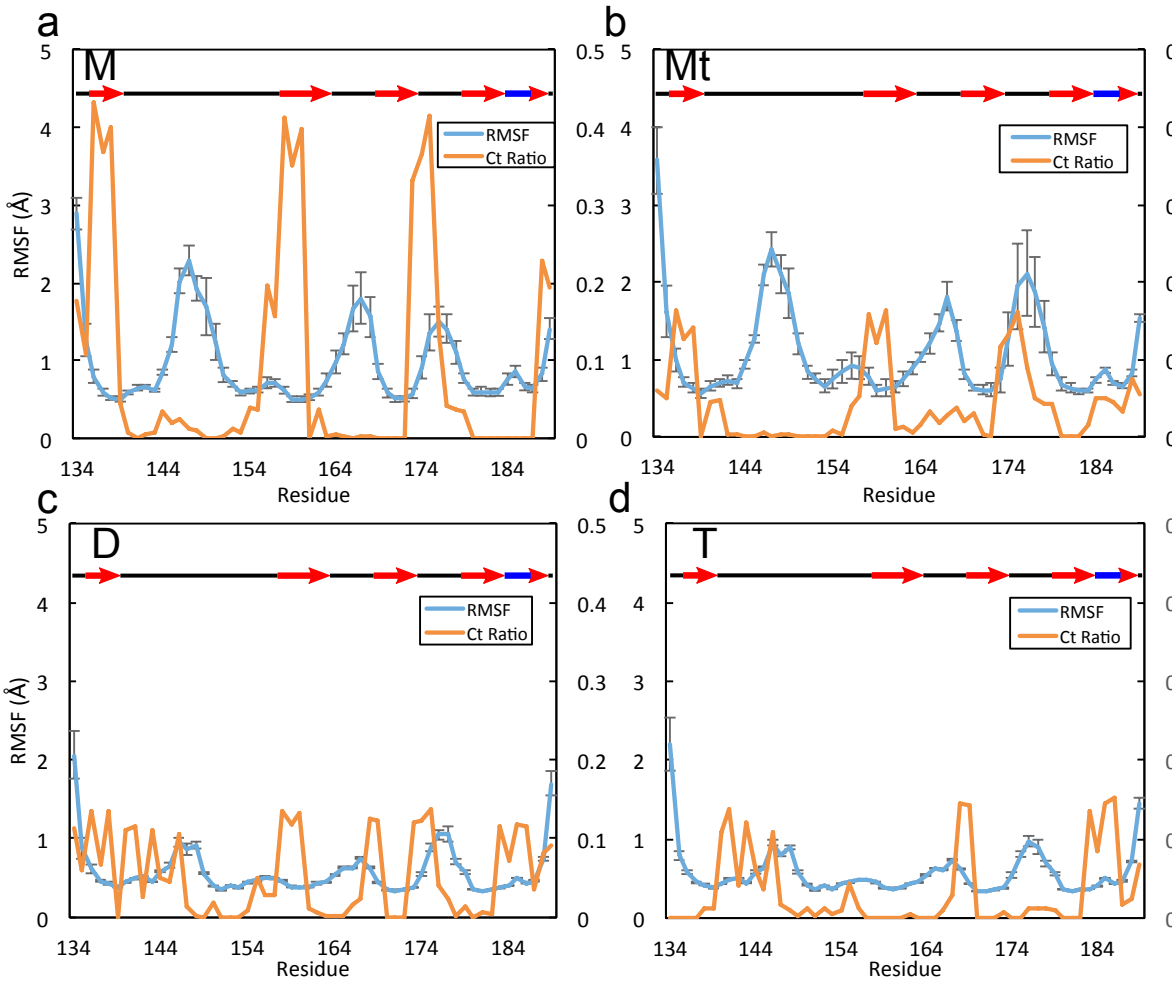


Figure 4.S2: RMSF of SH3 domain in the (a) monomer M, (b) ternary Mt, (c) dimer D, and (d) tetramer T systems. Error bars shown on RMSF plots are standard error.

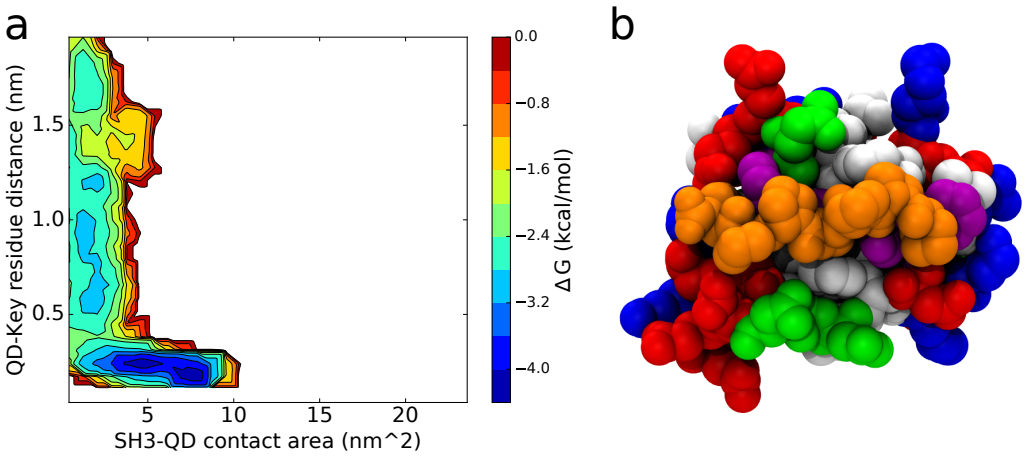


Figure 4.S3: (a) PRM-SH3 binding free energy surface. (b) Main binding well structure. PRM is shown in orange over the binding site residues (purple).

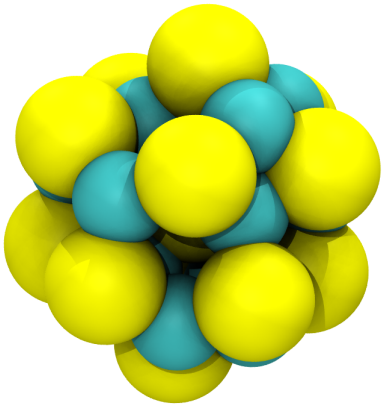


Figure 4.S4: $(CdSe)_{13}$ QD core.

Appendix

Appendix A: List of Abbrevations

AMOEBA: Atomic Multipole Optimized Energetics for Biomolecular Applications (a force field)

BAR: Bennett Acceptance Ratio

CB[n]: Cucurbituril macrocycle with n glycoluril subunits

CdSe: Cadmium Selenide

CG: Coarse grained

Cucurbit[7]uril: See CB[n]

MD: Molecular Dynamics

OSRW: Orthogonal Space Random Walk

PDB: Protein Data Bank

PMF: Potential of Mean Force

PRM: Proline-Rich Motif

QDs: Quantum Dots

RACER: RNA CoarsE gRained model

RMSD: Root Mean Square Deviation

RMSF: Root Mean Square Fluctuation

SH3: Src Homology 3 protein domain

TOPO: trioctylphosphine oxide

vdW_{eff}: Effective van der Waals interaction

WHAM: Weighted Histogram Analysis Method

References

1 Lai, D., Proctor, J. R. & Meyer, I. M. On the importance of cotranscriptional RNA structure formation. *RNA-Publ. RNA Soc.* **19**, 1461-1473, doi:10.1261/rna.037390.112 (2013).

2 Chauhan, S. & Woodson, S. A. Tertiary interactions determine the accuracy of RNA folding. *J. Am. Chem. Soc.* **130**, 1296-1303, doi:10.1021/ja076166i (2008).

3 Rangan, P., Masquida, B., Westhof, E. & Woodson, S. A. Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proceedings of the National Academy of Sciences* **100**, 1574-1579, doi:10.1073/pnas.0337743100 (2003).

4 Woodson, S. A. in *Annual Review of Biophysics, Vol 39* Vol. 39 *Annual Review of Biophysics* (eds D. C. Rees, K. A. Dill, & J. R. Williamson) 61-77 (Annual Reviews, 2010).

5 Mitchell, D. & Russell, R. Folding Pathways of the Tetrahymena Ribozyme. *J. Mol. Biol.* **426**, 2300-2312, doi:10.1016/j.jmb.2014.04.011 (2014).

6 Uhlenbeck, O. C. KEEPING RNA HAPPY. *RNA-Publ. RNA Soc.* **1**, 4-6 (1995).

7 Uhlenbeck, O. C. RNA biophysics has come of age. *Biopolymers* **91**, 811-814, doi:10.1002/bip.21269 (2009).

8 Karplus, M. The Levinthal paradox: yesterday and today. *Fold. Des.* **2**, S69-S75, doi:10.1016/s1359-0278(97)00067-9 (1997).

9 Zheng, H., Shabalin, I. G., Handing, K. B., Bujnicki, J. M. & Minor, W. Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Research* **43**, 3789-3801, doi:10.1093/nar/gkv225 (2015).

10 Doshi, K. J., Cannone, J. J., Cobaugh, C. W. & Gutell, R. R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5**, 1-22, doi:10.1186/1471-2105-5-105 (2004).

11 Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research* **38**, D280-D282, doi:10.1093/nar/gkp892 (2010).

12 Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51-55, doi:http://www.nature.com/nature/journal/v452/n7183/suppinfo/nature06684_S1.html (2008).

13 Cruz, J. A. *et al.* RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *Rna* **18**, 610-625, doi:10.1261/rna.031054.111 (2012).

14 Miao, Z. *et al*. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *Rna* **21**, 1066-1084, doi:10.1261/rna.049502.114 (2015).
15 Fera, D. *et al*. RAG: RNA-As-Graphs web resource. *BMC Bioinformatics* **5**, 1-9, doi:10.1186/1471-2105-5-88 (2004).
16 Izzo, J. A., Kim, N., Elmetwaly, S. & Schlick, T. RAG: An update to the RNA-As-Graphs resource. *Bmc Bioinformatics* **12**, 17, doi:10.1186/1471-2105-12-219 (2011).
17 Kim, N. *et al*. Graph-based sampling for approximating global helical topologies of RNA. *Proceedings of the National Academy of Sciences* **111**, 4079-4084, doi:10.1073/pnas.1318893111 (2014).
18 Kim, N., Petingi, L. & Schlick, T. Network Theory Tools for RNA Modeling. *WSEAS transactions on mathematics* **9**, 941-955 (2013).
19 Kim, N., Zahran, M. & Schlick, T. Computational prediction of riboswitch tertiary structures including pseudoknots by RAGTOP: a hierarchical graph sampling approach. *Methods in enzymology* **553**, 115-135, doi:10.1016/bs.mie.2014.10.054 (2015).
20 Zahran, M., Sevim Bayrak, C., Elmetwaly, S. & Schlick, T. RAG-3D: a search tool for RNA 3D substructures. *Nucleic Acids Research*, doi:10.1093/nar/gkv823 (2015).
21 Fulle, S. & Gohlke, H. Statics of the Ribosomal Exit Tunnel: Implications for Cotranslational Peptide Folding, Elongation Regulation, and Antibiotics Binding. *J. Mol. Biol.* **387**, 502-517, doi:http://dx.doi.org/10.1016/j.jmb.2009.01.037 (2009).
22 Gillespie, J., Mayne, M. & Jiang, M. RNA folding on the 3D triangular lattice. *BMC Bioinformatics* **10**, 1-17, doi:10.1186/1471-2105-10-369 (2009).
23 Kerpedjiev, P., Höner zu Siederdissen, C. & Hofacker, I. L. Predicting RNA 3D structure using a coarse-grain helix-centered model. *Rna* **21**, 1110-1121, doi:10.1261/rna.047522.114 (2015).
24 Lamiable, A., Quessette, F., Vial, S., Barth, D. & Denise, A. An Algorithmic Game-Theory Approach for Coarse-Grain Prediction of RNA 3D Structure. *Ieee-Acm Transactions on Computational Biology and Bioinformatics* **10**, 193-199, doi:10.1109/tcbb.2012.148 (2013).
25 Xia, Z., Gardner, D. P., Gutell, R. R. & Ren, P. Y. Coarse-Grained Model for Simulation of RNA Three-Dimensional Structures. *J. Phys. Chem. B* **114**, 13497-13506, doi:10.1021/jp104926t (2010).
26 Xia, Z. & Ren, P. in *Biophysics of RNA Folding* Vol. 3 *Biophysics for the Life Sciences* (ed Rick Russell) Ch. 4, 53-68 (Springer New York, 2013).
27 Cech, T. R., Zaug, A. J. & Grabowski, P. J. In vitro splicing of the ribosomal RNA precursor of tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**, 487-496, doi:http://dx.doi.org/10.1016/0092-8674(81)90390-1 (1981).

28 Kruger, K. *et al*. SELF-SPLICING RNA - AUTO-EXCISION AND AUTO-CYCLIZATION OF THE RIBOSOMAL-RNA INTERVENING SEQUENCE OF TETRAHYMENA. *Cell* **31**, 147-157, doi:10.1016/0092-8674(82)90414-7 (1982).
29 Guerriertakada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. THE RNA MOIETY OF RIBONUCLEASE-P IS THE CATALYTIC SUBUNIT OF THE ENZYME. *Cell* **35**, 849-857, doi:10.1016/0092-8674(83)90117-4 (1983).
30 Mironov, A. S. *et al*. Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria. *Cell* **111**, 747-756, doi:10.1016/s0092-8674(02)01134-0 (2002).
31 Nahvi, A. *et al*. Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9**, 1043-1049, doi:10.1016/s1074-5521(02)00224-7 (2002).
32 Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952-956, doi:10.1038/nature01145 (2002).
33 Breaker, R. R. Prospects for Riboswitch Discovery and Analysis. *Mol. Cell* **43**, 867-879, doi:10.1016/j.molcel.2011.08.024 (2011).
34 Serganov, A. & Nudler, E. A Decade of Riboswitches. *Cell* **152**, 17-24, doi:10.1016/j.cell.2012.12.024 (2013).
35 Russell, R. in *Biophysics of RNA Folding Biophysics for the Life Sciences* (ed R. Russell) Ch. 1, 1-10 (Springer-Verlag New York, 2013).
36 Mitchell, D., Jarmoskaite, I., Seval, N., Seifert, S. & Russell, R. The Long-Range P3 Helix of the Tetrahymena Ribozyme Is Disrupted during Folding between the Native and Misfolded Conformations. *J. Mol. Biol.* **425**, 2670-2686, doi:10.1016/j.jmb.2013.05.008 (2013).
37 Russell, R. *et al*. The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. *J. Mol. Biol.* **363**, 531-544, doi:10.1016/j.jmb.2006.08.024 (2006).
38 Russell, R. *et al*. Exploring the folding landscape of a structured RNA. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 155-160, doi:10.1073/pnas.221593598 (2002).
39 Thirumalai, D. & Hyeon, C. in *Non-Protein Coding RNAs* (eds Nils G. Walter, Sarah A. Woodson, & Robert T. Batey) 27-47 (Springer Berlin Heidelberg, 2009).
40 Silverman, S. K., Deras, M. L., Woodson, S. A., Scaringe, S. A. & Cech, T. R. Multiple Folding Pathways for the P4–P6 RNA Domain. *Biochemistry* **39**, 12465-12475, doi:10.1021/bi000828y (2000).
41 Woodson, S. A. Recent insights on RNA folding mechanisms from catalytic RNA. *Cell. Mol. Life Sci.* **57**, 796-808, doi:10.1007/s000180050042 (2000).
42 Schroeder, R., Barta, A. & Semrad, K. Strategies for RNA folding and assembly. *Nature Reviews Molecular Cell Biology* **5**, 908-919, doi:10.1038/nrm1497 (2004).
43 Bokinsky, G. & Zhuang, X. W. Single-molecule RNA folding. *Accounts Chem. Res.* **38**, 566-573, doi:10.1021/ar040142o (2005).

44 Gell, C. *et al*. Single-Molecule Fluorescence Resonance Energy Transfer Assays Reveal Heterogeneous Folding Ensembles in a Simple RNA Stem-Loop. *J. Mol. Biol*. **384**, 264-278, doi:10.1016/j.jmb.2008.08.088 (2008).

45 Schuster, P. Prediction of RNA secondary structures: from theory to models and real molecules. *Rep. Prog. Phys*. **69**, 1419-1477, doi:10.1088/0034-4885/69/5/r04 (2006).

46 Cannone, J. J. *et al*. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *Bmc Bioinformatics* **3**, doi:10.1186/1471-2105-3-2 (2002).

47 Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. & Stadler, P. F. RNAalifold: improved consensus structure prediction for RNA alignments. *Bmc Bioinformatics* **9**, 13, doi:10.1186/1471-2105-9-474 (2008).

48 Hofacker, I. L., Fekete, M. & Stadler, P. F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol*. **319**, 1059-1066, doi:10.1016/s0022-2836(02)00308-x (2002).

49 Knudsen, B. & Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research* **31**, 3423-3428, doi:10.1093/nar/gkg614 (2003).

50 Markham, N. & Zuker, M. in *Bioinformatics* Vol. 453 *Methods in Molecular Biology™* (ed JonathanM Keith) Ch. 1, 3-31 (Humana Press, 2008).

51 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* **31**, 3406-3415, doi:10.1093/nar/gkg595 (2003).

52 Hofacker, I. L. *et al*. FAST FOLDING AND COMPARISON OF RNA SECONDARY STRUCTURES. *Mon. Chem*. **125**, 167-188, doi:10.1007/bf00818163 (1994).

53 Hofacker, I. in *Comparative Genomics* Vol. 395 *Methods in Molecular Biology™* (ed NicholasH Bergman) Ch. 33, 527-543 (Humana Press, 2008).

54 Lorenz, R. *et al*. ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 1-14, doi:10.1186/1748-7188-6-26 (2011).

55 Bellaousov, S. & Mathews, D. H. ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *Rna* **16**, 1870-1880, doi:10.1261/rna.2125310 (2010).

56 Ren, J., Rastegari, B., Condon, A. & Hoos, H. H. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *Rna* **11**, 1494-1504, doi:10.1261/rna.7284905 (2005).

57 Sato, K., Kato, Y., Hamada, M., Akutsu, T. & Asai, K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* **27**, i85-i93, doi:10.1093/bioinformatics/btr215 (2011).

58 Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2[prime]-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protocols* **1**, 1610-1616, doi:10.1038/nprot.2006.249 (2006).

59 Lusvarghi, S., Sztuba-Solinska, J., Purzycka, K. J., Rausch, J. W. & Le Grice, S. F. J. RNA Secondary Structure Prediction Using High-throughput SHAPE. e50243, doi:doi:10.3791/50243 (2013).

60 Leonard, C. W. *et al*. Principles for Understanding the Accuracy of SHAPE-Directed RNA Structure Modeling. *Biochemistry* **52**, 588-595, doi:10.1021/bi300755u (2013).

61 Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. Understanding the Errors of SHAPE-Directed RNA Structure Modeling. *Biochemistry* **50**, 8049-8056, doi:10.1021/bi200524n (2011).

62 Sükösd, Z., Swenson, M. S., Kjems, J. & Heitsch, C. E. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Research* **41**, 2807-2816, doi:10.1093/nar/gks1283 (2013).

63 Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F. & Wolfinger, M. T. SHAPE directed RNA folding. *Bioinformatics* **32**, 145-147, doi:10.1093/bioinformatics/btv523 (2016).

64 Hajdin, C. E. *et al*. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences* **110**, 5498-5503, doi:10.1073/pnas.1219988110 (2013).

65 Laing, C. & Schlick, T. Computational approaches to RNA structure prediction, analysis, and design. *Current Opinion in Structural Biology* **21**, 306-318, doi:http://dx.doi.org/10.1016/j.sbi.2011.03.015 (2011).

66 Laing, C. & Schlick, T. Computational approaches to 3D modeling of RNA. *J. Phys.-Condes. Matter* **22**, 18, doi:10.1088/0953-8984/22/28/283101 (2010).

67 Frellsen, J. *et al*. A Probabilistic Model of RNA Conformational Space. *Plos Computational Biology* **5**, 11, doi:10.1371/journal.pcbi.1000406 (2009).

68 Bida, J. P. & Maher, L. J. Improved prediction of RNA tertiary structure with insights into native state dynamics. *RNA-Publ. RNA Soc.* **18**, 385-393, doi:10.1261/rna.027201.111 (2012).

69 Zhao, Y. J. *et al*. Automated and fast building of three-dimensional RNA structures. *Sci Rep* **2**, 6, doi:10.1038/srep00734 (2012).

70 Popenda, M. *et al*. Automated 3D structure composition for large RNAs. *Nucleic Acids Research* **40**, 12, doi:10.1093/nar/gks339 (2012).

71 Cao, S. & Chen, S.-J. Predicting RNA folding thermodynamics with a reduced chain representation model. *Rna* **11**, 1884-1897, doi:10.1261/rna.2109105 (2005).

72 Cao, S. & Chen, S. J. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA-Publ. RNA Soc.* **15**, 696-706, doi:10.1261/rna.1429009 (2009).

73 Cao, S. & Chen, S. J. Physics-Based De Novo Prediction of RNA 3D Structures. *J. Phys. Chem. B* **115**, 4216-4226, doi:10.1021/jp112059y (2011).

74 Xu, X. J., Zhao, P. N. & Chen, S. J. Vfold: A Web Server for RNA Structure and Folding Thermodynamics Prediction. *PLoS One* **9**, 7, doi:10.1371/journal.pone.0107504 (2014).

75 Reinharz, V., Major, F. & Waldispühl, J. Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics* **28**, i207-i214, doi:10.1093/bioinformatics/bts226 (2012).
76 Das, R. & Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences* **104**, 14664-14669, doi:10.1073/pnas.0703836104 (2007).
77 Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods* **7**, 291-294, doi:10.1038/nmeth.1433 (2010).
78 Cheng, C. Y., Chou, F.-C. & Das, R. in *Methods in Enzymology* Vol. Volume 553 (eds Chen Shi-Jie & H. Burke-Aguero Donald) 35-64 (Academic Press, 2015).
79 Leaver-Fay, A. *et al*. in *Methods in Enzymology* Vol. Volume 487 (eds L. Johnson Michael & Brand Ludwig) 545-574 (Academic Press, 2011).
80 Jossinet, F., Ludwig, T. E. & Westhof, E. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics* **26**, 2057-2059, doi:10.1093/bioinformatics/btq321 (2010).
81 Martinez, H. M., Maizel, J. V. & Shapiro, B. A. RNA2D3D: A program for Generating, Viewing, and Comparing 3-Dimensional Models of RNA. *Journal of Biomolecular Structure and Dynamics* **25**, 669-683, doi:10.1080/07391102.2008.10531240 (2008).
82 Dawson, W. K., Maciejczyk, M., Jankowska, E. J. & Bujnicki, J. M. Coarse-grained modeling of RNA 3D structure. *Methods*, doi:http://dx.doi.org/10.1016/j.ymeth.2016.04.026.
83 Malhotra, A., Tan, R. K. Z. & Harvey, S. C. MODELING LARGE RNAS AND RIBONUCLEOPROTEIN-PARTICLES USING MOLECULAR MECHANICS TECHNIQUES. *Biophys. J.* **66**, 1777-1795 (1994).
84 Tan, R. K. Z., Petrov, A. S. & Harvey, S. C. YUP: A molecular simulation program for coarse-grained and multiscaled models. *Journal of Chemical Theory and Computation* **2**, 529-540, doi:10.1021/ct050323r (2006).
85 Jonikas, M. A., Radmer, R. J. & Altman, R. B. Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics* **25**, 3259-3266, doi:10.1093/bioinformatics/btp576 (2009).
86 Jonikas, M. A. *et al*. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *Rna* **15**, 189-199, doi:10.1261/rna.1270809 (2009).
87 Krokhotin, A., Houlihan, K. & Dokholyan, N. V. iFoldRNA v2: folding RNA with constraints. *Bioinformatics*, doi:10.1093/bioinformatics/btv221 (2015).
88 Sharma, S., Ding, F. & Dokholyan, N. V. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* **24**, 1951-1952, doi:10.1093/bioinformatics/btn328 (2008).

89 Denesyuk, N. A. & Thirumalai, D. Coarse-Grained Model for Predicting RNA Folding Thermodynamics. *J. Phys. Chem. B* **117**, 4901-4911, doi:10.1021/jp401087x (2013).

90 Denesyuk, N. A. & ThirumalaiD. How do metal ions direct ribozyme folding? *Nat Chem* **7**, 793-801, doi:10.1038/nchem.2330 http://www.nature.com/nchem/journal/v7/n10/abs/nchem.2330.html - supplementary-information (2015).

91 Mustoe, A. M., Al-Hashimi, H. M. & Brooks, C. L. Coarse Grained Models Reveal Essential Contributions of Topological Constraints to the Conformational Free Energy of RNA Bulges. *The Journal of Physical Chemistry B* **118**, 2615-2627, doi:10.1021/jp411478x (2014).

92 Mustoe, A. M., Brooks, C. L. & Al-Hashimi, H. M. Topological constraints are major determinants of tRNA tertiary structure and dynamics and provide basis for tertiary folding cooperativity. *Nucleic Acids Research* **42**, 11792-11804, doi:10.1093/nar/gku807 (2014).

93 Mustoe, A. M. *et al*. Noncanonical Secondary Structure Stabilizes Mitochondrial tRNASer(UCN) by Reducing the Entropic Cost of Tertiary Folding. *J. Am. Chem. Soc*. **137**, 3592-3599, doi:10.1021/ja5130308 (2015).

94 Cragnolini, T., Derreumaux, P. & Pasquali, S. Coarse-Grained Simulations of RNA and DNA Duplexes. *J. Phys. Chem. B* **117**, 8047-8060, doi:10.1021/jp400786b (2013).

95 Pasquali, S. & Derreumaux, P. HiRE-RNA: A High Resolution Coarse-Grained Energy Model for RNA. *The Journal of Physical Chemistry B* **114**, 11957-11966, doi:10.1021/jp102497y (2010).

96 Cragnolini, T., Laurin, Y., Derreumaux, P. & Pasquali, S. Coarse-Grained HiRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings. *Journal of Chemical Theory and Computation* **11**, 3510-3522, doi:10.1021/acs.jctc.5b00200 (2015).

97 Boniecki, M. J. *et al*. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research* **44**, e63, doi:10.1093/nar/gkv1479 (2016).

98 Magnus, M., Boniecki, M. J., Dawson, W. & Bujnicki, J. M. SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Research*, doi:10.1093/nar/gkw279 (2016).

99 Bernauer, J., Huang, X., Sim, A. Y. L. & Levitt, M. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *Rna* **17**, 1066-1075, doi:10.1261/rna.2543711 (2011).

100 Xia, Z., Bell, D. R., Shi, Y. & Ren, P. RNA 3D Structure Prediction by Using a Coarse-Grained Model and Experimental Data. *The Journal of Physical Chemistry B* **117**, 3135-3144, doi:10.1021/jp400751w (2013).

101 TINKER Molecular Modeling Package v. 6.3 (http://dasher.wustl.edu/tinker).

102 Wang, L.-P., Chen, J. & Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *Journal of Chemical Theory and Computation* **9**, 452-460, doi:10.1021/ct300826t (2013).
103 Hyeon, C., Dima, R. I. & Thirumalai, D. Size, shape, and flexibility of RNA structures. *The Journal of Chemical Physics* **125**, 194905, doi:doi:http://dx.doi.org/10.1063/1.2364190 (2006).
104 Saunders, M. G. & Voth, G. A. Coarse-Graining Methods for Computational Biology. *Annual Review of Biophysics* **42**, 73-93, doi:10.1146/annurev-biophys-083012-130348 (2013).
105 Müller-Plathe, F. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **3**, 754-769, doi:10.1002/1439-7641(20020916)3:9<754::AID-CPHC754>3.0.CO;2-U (2002).
106 Tschöp, W., Kremer, K., Batoulis, J., Bürger, T. & Hahn, O. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polymerica* **49**, 61-74, doi:10.1002/(SICI)1521-4044(199802)49:2/3<61::AID-APOL61>3.0.CO;2-V (1998).
107 Zhao, F. & Xu, J. A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study. *Structure* **20**, 1118-1126, doi:http://dx.doi.org/10.1016/j.str.2012.04.003 (2012).
108 Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science* **11**, 2714-2726, doi:10.1110/ps.0217002 (2002).
109 Shen, M.-y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Science : A Publication of the Protein Society* **15**, 2507-2524, doi:10.1110/ps.062416606 (2006).
110 Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223-230, doi:10.1126/science.181.4096.223 (1973).
111 Yakovchuk, P., Protozanova, E. & Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research* **34**, 564-574, doi:10.1093/nar/gkj454 (2006).
112 Xia, T. B. *et al*. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719-14735, doi:10.1021/bi9809425 (1998).
113 Mathews, D. H. *et al*. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7287-7292, doi:10.1073/pnas.0401799101 (2004).
114 Freier, S. M. *et al*. IMPROVED FREE-ENERGY PARAMETERS FOR PREDICTIONS OF RNA DUPLEX STABILITY. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 9373-9377, doi:10.1073/pnas.83.24.9373 (1986).

115 Borer, P. N., Dengler, B., Tinoco Jr, I. & Uhlenbeck, O. C. Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol.* **86**, 843-853, doi:http://dx.doi.org/10.1016/0022-2836(74)90357-X (1974).
116 Breslauer, K. J., Frank, R., Blocker, H. & Marky, L. A. PREDICTING DNA DUPLEX STABILITY FROM THE BASE SEQUENCE. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 3746-3750, doi:10.1073/pnas.83.11.3746 (1986).
117 Xia, T. B., McDowell, J. A. & Turner, D. H. Thermodynamics of nonsymmetric tandem mismatches adjacent to G center dot C base pairs in RNA. *Biochemistry* **36**, 12486-12497, doi:10.1021/bi971069v (1997).
118 Li, P. T. X., Collin, D., Smith, S. B., Bustamante, C. & Tinoco, I. Probing the mechanical folding kinetics of TAR RNA by hopping, force-jump, and force-ramp methods. *Biophys. J.* **90**, 250-260, doi:10.1529/biophysj.105.068049 (2006).
119 WHAM: The Weighted Histogram Analysis Method v. 2.0.9 (http://membrane.urmc.rochester.edu/content/wham).
120 Sharp, K. A. in *Protein-Ligand Interactions* 1-22 (Wiley-VCH Verlag GmbH & Co. KGaA, 2012).
121 Dale, T., Smith, R. & Serra, M. J. A test of the model to predict unusually stable RNA hairpin loop stability. *Rna* **6**, 608-615 (2000).
122 Giese, M. R. *et al.* Stability of RNA Hairpins Closed by Wobble Base Pairs. *Biochemistry* **37**, 1094-1100, doi:10.1021/bi972050v (1998).
123 Groebe, D. R. & Uhlenbeck, O. C. CHARACTERIZATION OF RNA HAIRPIN LOOP STABILITY. *Nucleic Acids Research* **16**, 11725-11735, doi:10.1093/nar/16.24.11725 (1988).
124 Antao, V. P. & Tinoco, I. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Research* **20**, 819-824, doi:10.1093/nar/20.4.819 (1992).
125 Serra, M. J., Lyttle, M. H., Axenson, T. J., Schadt, C. A. & Turner, D. H. RNA hairpin loop stability depends on closing base pair. *Nucleic Acids Research* **21**, 3845-3849, doi:10.1093/nar/21.16.3845 (1993).
126 Burkard, M. E., Kierzek, R. & Turner, D. H. Thermodynamics of unpaired terminal nucleotides on short RNA helixes correlates with stacking at helix termini in larger RNAs1. *J. Mol. Biol.* **290**, 967-982, doi:http://dx.doi.org/10.1006/jmbi.1999.2906 (1999).
127 Woodside, M. T. *et al.* Direct Measurement of the Full, Sequence-Dependent Folding Landscape of a Nucleic Acid. *Science* **314**, 1001-1004, doi:10.1126/science.1133601 (2006).
128 Woodside, M. T. *et al.* Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins. *Proceedings of the National Academy of Sciences* **103**, 6190-6195, doi:10.1073/pnas.0511048103 (2006).
129 Liphardt, J., Onoa, B., Smith, S. B., Tinoco, I. & Bustamante, C. Reversible Unfolding of Single RNA Molecules by Mechanical Force. *Science* **292**, 733-737, doi:10.1126/science.1058498 (2001).

130 Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**, 1011-1021, doi:10.1002/jcc.540130812 (1992).
131 Eastman, P. *et al.* OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation* **9**, 461-469, doi:10.1021/ct300857j (2013).
132 Malmberg, C. G. & Maryott, A. A. DIELECTRIC CONSTANT OF WATER FROM 0-DEGREES-C TO 100-DEGREES-C. *J. Res. Natl. Bur. Stand.* **56**, 1-8, doi:10.6028/jres.056.001 (1956).
133 Gilson, M. K. & Honig, B. H. THE DIELECTRIC-CONSTANT OF A FOLDED PROTEIN. *Biopolymers* **25**, 2097-2119, doi:10.1002/bip.360251106 (1986).
134 Schutz, C. N. & Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins-Structure Function and Bioinformatics* **44**, 400-417, doi:10.1002/prot.1106 (2001).
135 Israelachvili, J. N. in *Intermolecular and Surface Forces (Third Edition)* 291-340 (Academic Press, 2011).
136 Conte, M. R., Conn, G. L., Brown, T. & Lane, A. N. Conformational properties and thermodynamics of the RNA duplex r(CGCAAAUUUGCG)2: comparison with the DNA analogue d(CGCAAATTTGCG)2. *Nucleic Acids Research* **25**, 2627-2634, doi:10.1093/nar/25.13.2627 (1997).
137 Lii, J.-H. & Allinger, N. L. Directional hydrogen bonding in the MM3 force field: II. *J. Comput. Chem.* **19**, 1001-1016, doi:10.1002/(SICI)1096-987X(19980715)19:9<1001::AID-JCC2>3.0.CO;2-U (1998).
138 Persch, E., Dumele, O. & Diederich, F. Molecular Recognition in Chemical and Biological Systems. *Angew. Chem.-Int. Edit.* **54**, 3290-3327, doi:10.1002/anie.201408487 (2015).
139 Gohlke, H. & Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie (International ed. in English)* **41**, 2644-2676, doi:10.1002/1521-3773(20020802)41:15<2644::aid-anie2644>3.0.co;2-o (2002).
140 Houk, K. N., Leach, A. G., Kim, S. P. & Zhang, X. Y. Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. *Angew. Chem.-Int. Edit.* **42**, 4872-4897, doi:10.1002/anie.200200565 (2003).
141 Gilson, M. K. & Zhou, H. X. in *Annual Review of Biophysics and Biomolecular Structure* Vol. 36 *Annual Review of Biophysics* 21-42 (Annual Reviews, 2007).
142 Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**, 935-949, doi:http://www.nature.com/nrd/journal/v3/n11/suppinfo/nrd1549_S1.html (2004).
143 Grater, F., Schwarzl, S. M., Dejaegere, A., Fischer, S. & Smith, J. C. Protein/ligand binding free energies calculated with quantum

mechanics/molecular mechanics. *J. Phys. Chem. B* **109**, 10474-10483, doi:10.1021/jp044185y (2005).

144 Mikulskis, P. *et al*. Free-energy perturbation and quantum mechanical study of SAMPL4 octa-acid host-guest binding energies. *Journal of Computer-Aided Molecular Design* **28**, 375-400, doi:10.1007/s10822-014-9739-x (2014).

145 Anisimov, V. M. & Cavasotto, C. N. Quantum Mechanical Binding Free Energy Calculation for Phosphopeptide Inhibitors of the Lck SH2 Domain. *J. Comput. Chem*. **32**, 2254-2263, doi:10.1002/jcc.21808 (2011).

146 Lawrenz, M., Wereszczynski, J., Ortiz-Sanchez, J. M., Nichols, S. E. & McCammon, J. A. Thermodynamic integration to predict host-guest binding affinities. *J. Comput.-Aided Mol. Des*. **26**, 569-576, doi:10.1007/s10822-012-9542-5 (2012).

147 Monroe, J. I. & Shirts, M. R. Converging free energies of binding in cucurbit 7 uril and octa-acid host-guest systems from SAMPL4 using expanded ensemble simulations. *Journal of Computer-Aided Molecular Design* **28**, 401-415, doi:10.1007/s10822-014-9716-4 (2014).

148 D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman. (University of California, San Francisco, 2015).

149 Vanommeslaeghe, K. *et al*. CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem*. **31**, 671-690, doi:10.1002/jcc.21367 (2010).

150 Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of Chemical Theory and Computation* **11**, 3499-3509, doi:10.1021/acs.jctc.5b00356 (2015).

151 Reif, M. M., Hünenberger, P. H. & Oostenbrink, C. New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *Journal of Chemical Theory and Computation* **8**, 3705-3723, doi:10.1021/ct300156h (2012).

152 Ponder, J. W. & Case, D. A. Force fields for protein simulations. *Adv.Protein Chem*. **66**, 27-+ (2003).

153 Rick, S. W. & Stuart, S. J. in *Reviews in Computational Chemistry, Vol 18* Vol. 18 *Reviews in Computational Chemistry* (eds K. B. Lipkowitz & D. B. Boyd) 89-146 (Wiley-Vch, Inc, 2002).

154 Williams, D. E. REPRESENTATION OF THE MOLECULAR ELECTROSTATIC POTENTIAL BY ATOMIC MULTIPOLE AND BOND DIPOLE MODELS. *Journal of Computational Chemistry* **9**, 745-763, doi:10.1002/jcc.540090705 (1988).

155 Ren, P. & Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *The Journal of Physical Chemistry B* **107**, 5933-5947, doi:10.1021/jp027815+ (2003).
156 Patel, S. & Brooks, C. L. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* **25**, 1-16, doi:10.1002/jcc.10355 (2004).
157 Baker, C. M., Anisimov, V. M. & MacKerell, A. D. Development of CHARMM Polarizable Force Field for Nucleic Acid Bases Based on the Classical Drude Oscillator Model. *The Journal of Physical Chemistry B* **115**, 580-596, doi:10.1021/jp1092338 (2011).
158 Lemkul, J. A., Huang, J., Roux, B. & MacKerell, A. D. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chemical Reviews*, doi:10.1021/acs.chemrev.5b00505 (2016).
159 Shi, Y. *et al*. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation* **9**, 4046-4063, doi:10.1021/ct4003702 (2013).
160 Ren, P., Wu, C. & Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **7**, 3143-3161, doi:10.1021/ct200304d (2011).
161 Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National Academy of Sciences* **112**, E386-E391, doi:10.1073/pnas.1424461112 (2015).
162 Gumbart, J. C., Roux, B. & Chipot, C. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* **9**, 794-802, doi:10.1021/ct3008099 (2013).
163 Bennett, C. H. EFFICIENT ESTIMATION OF FREE-ENERGY DIFFERENCES FROM MONTE-CARLO DATA. *J. Comput. Phys.* **22**, 245-268, doi:10.1016/0021-9991(76)90078-4 (1976).
164 Christ, C. D., Mark, A. E. & van Gunsteren, W. F. Basic ingredients of free energy calculations: A review. *Journal of Computational Chemistry* **31**, 1569-1582, doi:10.1002/jcc.21450 (2010).
165 Daniel, M. Z. Equilibrium Sampling in Biomolecular Simulations. *Annual Review of Biophysics* **40**, 41-62, doi:doi:10.1146/annurev-biophys-042910-155255 (2011).
166 Muddana, H. S. *et al*. Blind prediction of host-guest binding affinities: a new SAMPL3 challenge. *Journal of Computer-Aided Molecular Design* **26**, 475-487, doi:10.1007/s10822-012-9554-1 (2012).
167 Muddana, H. S., Fenley, A. T., Mobley, D. L. & Gilson, M. K. The SAMPL4 host-guest blind prediction challenge: an overview. *Journal of Computer-Aided Molecular Design* **28**, 305-317, doi:10.1007/s10822-014-9735-1 (2014).

168 Masson, E., Ling, X. X., Joseph, R., Kyeremeh-Mensah, L. & Lu, X. Y. Cucurbituril chemistry: a tale of supramolecular success. *RSC Adv*. **2**, 1213-1247, doi:10.1039/c1ra00768h (2012).
169 Walker, S., Oun, R., McInnes, F. J. & Wheate, N. J. The Potential of Cucurbit n urils in Drug Delivery. *Isr. J. Chem*. **51**, 616-624, doi:10.1002/ijch.201100033 (2011).
170 Lee, J. W., Samal, S., Selvapalam, N., Kim, H. J. & Kim, K. Cucurbituril homologues and derivatives: New opportunities in supramolecular chemistry. *Accounts Chem. Res*. **36**, 621-630, doi:10.1021/ar020254k (2003).
171 Jeon, Y. J. *et al*. Novel molecular drug carrier: encapsulation of oxaliplatin in cucurbit 7 uril and its effects on stability and reactivity of the drug. *Org. Biomol. Chem*. **3**, 2122-2125, doi:10.1039/b504487a (2005).
172 Ponder, J. W. *et al*. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **114**, 2549-2564, doi:10.1021/jp910674d (2010).
173 Mackerell, A. D. Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry* **25**, 1584-1604, doi:10.1002/jcc.20082 (2004).
174 Cieplak, P., Dupradeau, F. Y., Duan, Y. & Wang, J. M. Polarization effects in molecular mechanical force fields. *J Phys-Condens Mat* **21**, doi:Artn 333102
10.1088/0953-8984/21/33/333102 (2009).
175 Lopes, P. E. M., Roux, B. & MacKerell, A. D. Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications. *Theor Chem Acc* **124**, 11-28, doi:10.1007/s00214-009-0617-x (2009).
176 Marquez, C. & Nau, W. M. Polarizabilities inside molecular containers. *Angew. Chem.-Int. Edit*. **40**, 4387-+, doi:10.1002/1521-3773(20011203)40:23<4387::aid-anie4387>3.0.co;2-h (2001).
177 Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics* **129**, 124105, doi:doi:http://dx.doi.org/10.1063/1.2978177 (2008).
178 Zheng, L., Chen, M. & Yang, W. Simultaneous escaping of explicit and hidden free energy barriers: Application of the orthogonal space random walk strategy in generalized ensemble based conformational sampling. *J. Chem. Phys*. **130**, doi:10.1063/1.3153841 (2009).
179 Zheng, L., Chen, M. & Yang, W. Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 20227-20232, doi:10.1073/pnas.0810631106 (2008).
180 Zheng, L. & Yang, W. Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. *Journal of Chemical Theory and Computation* **8**, 810-823, doi:10.1021/ct200726v (2012).
181 Min, D. *et al*. Practically Efficient QM/MM Alchemical Free Energy Simulations: The Orthogonal Space Random Walk Strategy. *Journal of Chemical Theory and Computation* **6**, 2253-2266, doi:10.1021/ct100033s (2010).

182 Pearlman, D. A. & Kollman, P. A. The Lag between the Hamiltonian and the System Configuration in Free-Energy Perturbation Calculations. *Journal of Chemical Physics* **91**, 7831-7839, doi:Doi 10.1063/1.457251 (1989).
183 Kong, X. J. & Brooks, C. L. lambda-Dynamics: A new approach to free energy calculations. *Journal of Chemical Physics* **105**, 2414-2423, doi:Doi 10.1063/1.472109 (1996).
184 Abella, J. R., Cheng, S. Y., Wang, Q., Yang, W. & Ren, P. Hydration Free Energy from Orthogonal Space Random Walk and Polarizable Force Field. *Journal of Chemical Theory and Computation* **10**, 2792-2801, doi:10.1021/ct500202q (2014).
185 Jiao, D., Golubkov, P. A., Darden, T. A. & Ren, P. Calculation of protein-ligand binding free energy by using a polarizable potential. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6290-6295, doi:Doi 10.1073/Pnas.0711686105 (2008).
186 Hamelberg, D. & McCammon, J. A. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: Double-decoupling method. *J Am Chem Soc* **126**, 7683-7689, doi:Doi 10.1021/Ja0377908 (2004).
187 Jiao, D. *et al*. Trypsin-Ligand Binding Free Energies from Explicit and Implicit Solvent Simulations with Polarizable Potential. *Journal of Computational Chemistry* **30**, 1701-1711, doi:10.1002/jcc.21268 (2009).
188 Rocklin, G. J., Mobley, D. L., Dill, K. A. & Hünenberger, P. H. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *The Journal of Chemical Physics* **139**, 184103, doi:doi:http://dx.doi.org/10.1063/1.4826261 (2013).
189 Tuckerman, M., Berne, B. J. & Martyna, G. J. REVERSIBLE MULTIPLE TIME SCALE MOLECULAR-DYNAMICS. *J. Chem. Phys.* **97**, 1990-2001, doi:10.1063/1.463137 (1992).
190 Bussi, G., Donadio, D. & Parrinello, M. COMP 8-Canonical sampling through velocity rescaling. *Abstr Pap Am Chem S* **234** (2007).
191 Henriksen, N. M., Fenley, A. T. & Gilson, M. K. Computational Calorimetry: High-Precision Calculation of Host–Guest Binding Thermodynamics. *Journal of Chemical Theory and Computation* **11**, 4377-4394, doi:10.1021/acs.jctc.5b00405 (2015).
192 Wyczalkowski, M. A., Vitalis, A. & Pappu, R. V. New Estimators for Calculating Solvation Entropy and Enthalpy and Comparative Assessments of Their Accuracy and Precision. *The Journal of Physical Chemistry B* **114**, 8166-8180, doi:10.1021/jp103050u (2010).
193 Brooks, B. R., Janezic, D. & Karplus, M. HARMONIC-ANALYSIS OF LARGE SYSTEMS .1. METHODOLOGY. *J. Comput. Chem.* **16**, 1522-1542, doi:10.1002/jcc.540161209 (1995).

194 Andricioaei, I. & Karplus, M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* **115**, 6289-6292, doi:10.1063/1.1401821 (2001).
195 Chang, C. E., Chen, W. & Gilson, M. K. Evaluating the accuracy of the quasiharmonic approximation. *Journal of Chemical Theory and Computation* **1**, 1017-1028, doi:10.1021/ct0500904 (2005).
196 Baron, R., Hünenberger, P. H. & McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput.* **5**, 3150-3160, doi:10.1021/ct900373z (2009).
197 Wereszczynski, J. & McCammon, J. A. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Q. Rev. Biophys.* **45**, 1-25, doi:10.1017/s0033583511000096 (2012).
198 Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **100**, 020603 (2008).
199 Tu, Y. *et al.* Destructive extraction of phospholipids from Escherichia coli membranes by graphene nanosheets. *Nat Nano* **8**, 594-601, doi:10.1038/nnano.2013.125
http://www.nature.com/nnano/journal/v8/n8/abs/nnano.2013.125.html - supplementary-information (2013).
200 Zhou, R. & Gao, H. Cytotoxicity of graphene: recent advances and future perspective. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **6**, 452-474, doi:10.1002/wnan.1277 (2014).
201 Gao, J. *et al.* Size-dependent impact of CNTs on dynamic properties of calmodulin. *Nanoscale* **6**, 12828-12837, doi:10.1039/C4NR01623H (2014).
202 El-Sayed, R. *et al.* Single-Walled Carbon Nanotubes Inhibit the Cytochrome P450 Enzyme, CYP3A4. *Sci Rep* **6**, 21316, doi:10.1038/srep21316
http://www.nature.com/articles/srep21316 - supplementary-information (2016).
203 Zuo, G., Kang, S.-g., Xiu, P., Zhao, Y. & Zhou, R. Interactions Between Proteins and Carbon-Based Nanoparticles: Exploring the Origin of Nanotoxicity at the Molecular Level. *Small* **9**, 1546-1556, doi:10.1002/smll.201201381 (2013).
204 Ge, C. *et al.* Binding of blood proteins to carbon nanotubes reduces cytotoxicity. *Proceedings of the National Academy of Sciences* **108**, 16968-16973, doi:10.1073/pnas.1105270108 (2011).
205 Liu, Y., Zhao, Y., Sun, B. & Chen, C. Understanding the Toxicity of Carbon Nanotubes. *Accounts Chem. Res.* **46**, 702-713, doi:10.1021/ar300028m (2013).
206 Seabra, A. B., Paula, A. J., de Lima, R., Alves, O. L. & Durán, N. Nanotoxicity of Graphene and Graphene Oxide. *Chemical Research in Toxicology* **27**, 159-168, doi:10.1021/tx400385x (2014).
207 Johnston, H. J., Hutchison, G. R., Christensen, F. M., Aschberger, K. & Stone, V. The Biological Mechanisms and Physicochemical Characteristics Responsible for

Driving Fullerene Toxicity. *Toxicological Sciences* **114**, 162-182, doi:10.1093/toxsci/kfp265 (2010).
208 Aschberger, K. *et al*. Review of fullerene toxicity and exposure – Appraisal of a human health risk assessment, based on open literature. *Regulatory Toxicology and Pharmacology* **58**, 455-473, doi:http://dx.doi.org/10.1016/j.yrtph.2010.08.017 (2010).
209 Kang, S. G. *et al*. Dual Inhibitory Pathways of Metallofullerenol Gd@C-82(OH)(22) on Matrix Metalloproteinase-2: Molecular insight into drug-like nanomedicine. *Sci Rep* **4**, 8, doi:10.1038/srep04775 (2014).
210 Kang, S. G. *et al*. Molecular mechanism of pancreatic tumor metastasis inhibition by Gd@C-82(OH)(22) and its implication for de novo design of nanomedicine. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 15431-15436, doi:10.1073/pnas.1204600109 (2012).
211 Liang, X.-J. *et al*. Metallofullerene nanoparticles circumvent tumor resistance to cisplatin by reactivating endocytosis. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 7449-7454 (2010).
212 Kamat, P. V. Quantum Dot Solar Cells. The Next Big Thing in Photovoltaics. *The Journal of Physical Chemistry Letters* **4**, 908-918, doi:10.1021/jz400052e (2013).
213 Carey, G. H. *et al*. Colloidal Quantum Dot Solar Cells. *Chemical Reviews* **115**, 12732-12763, doi:10.1021/acs.chemrev.5b00063 (2015).
214 Jamieson, T. *et al*. Biological applications of quantum dots. *Biomaterials* **28**, 4717-4732, doi:http://dx.doi.org/10.1016/j.biomaterials.2007.07.014 (2007).
215 Yu, W. W., Chang, E., Drezek, R. & Colvin, V. L. Water-soluble quantum dots for biomedical applications. *Biochemical and Biophysical Research Communications* **348**, 781-786, doi:http://dx.doi.org/10.1016/j.bbrc.2006.07.160 (2006).
216 Rosenthal, S. J., Chang, J. C., Kovtun, O., McBride, J. R. & Tomlinson, I. D. Biocompatible Quantum Dots for Biological Applications. *Chem. Biol.* **18**, 10-24, doi:http://dx.doi.org/10.1016/j.chembiol.2010.11.013 (2011).
217 Wegner, K. D. & Hildebrandt, N. Quantum dots: bright and versatile in vitro and in vivo fluorescence imaging biosensors. *Chem. Soc. Rev.* **44**, 4792-4834, doi:10.1039/C4CS00532E (2015).
218 Sun, K. *et al*. Applications of colloidal quantum dots. *Microelectronics Journal* **40**, 644-649, doi:http://dx.doi.org/10.1016/j.mejo.2008.06.033 (2009).
219 Hardman, R. A Toxicologic Review of Quantum Dots: Toxicity Depends on Physicochemical and Environmental Factors. *Environmental Health Perspectives* **114**, 165-172 (2006).
220 Tsoi, K. M., Dai, Q., Alman, B. A. & Chan, W. C. W. Are Quantum Dots Toxic? Exploring the Discrepancy Between Cell Culture and Animal Studies. *Accounts Chem. Res.* **46**, 662-671, doi:10.1021/ar300040z (2013).
221 Stern, S. T. *et al*. Induction of Autophagy in Porcine Kidney Cells by Quantum Dots: A Common Cellular Response to Nanomaterials? *Toxicological Sciences* **106**, 140-152, doi:10.1093/toxsci/kfn137 (2008).

222 Pawson, T. & Nash, P. Assembly of Cell Regulatory Systems Through Protein Interaction Domains. *Science* **300**, 445-452, doi:10.1126/science.1083653 (2003).
223 Pawson, T. Specificity in Signal Transduction: From Phosphotyrosine-SH2 Domain Interactions to Complex Cellular Systems. *Cell* **116**, 191-203, doi:http://dx.doi.org/10.1016/S0092-8674(03)01077-8 (2004).
224 Pawson, T. & Nash, P. Protein–protein interactions define specificity in signal transduction. *Genes & Development* **14**, 1027-1047, doi:10.1101/gad.14.9.1027 (2000).
225 Scott, J. D. & Pawson, T. Cell Signaling in Space and Time: Where Proteins Come Together and When They're Apart. *Science* **326**, 1220-1224, doi:10.1126/science.1175668 (2009).
226 Mayer, B. J. SH3 domains: complexity in moderation. *Journal of Cell Science* **114**, 1253-1263 (2001).
227 Wu, X. *et al*. Structural basis for the specific interaction of lysine-containing proline-rich peptides with the N-terminal SH3 domain of c-Crk. *Structure* **3**, 215-226, doi:http://dx.doi.org/10.1016/S0969-2126(01)00151-4 (1995).
228 Best, R. B. *et al*. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi 1$ and $\chi 2$ Dihedral Angles. *Journal of Chemical Theory and Computation* **8**, 3257-3273, doi:10.1021/ct300400x (2012).
229 Schlenkrich, M., Brickmann, J., MacKerell, A. D. & Karplus, M. in *Biological Membranes: A Molecular Perspective from Computation and Experiment* (eds Kenneth M. Merz & Benoît Roux) 31-81 (Birkhäuser Boston, 1996).
230 Feller, S. E., Yin, D., Pastor, R. W. & MacKerell, A. D. Molecular dynamics simulation of unsaturated lipid bilayers at low hydration: parameterization and comparison with diffraction studies. *Biophys. J.* **73**, 2269-2279, doi:http://dx.doi.org/10.1016/S0006-3495(97)78259-6 (1997).
231 Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089-10092, doi:doi:http://dx.doi.org/10.1063/1.464397 (1993).
232 Phillips, J. C. *et al*. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781-1802, doi:10.1002/jcc.20289 (2005).
233 Kumar, S., Huang, C., Zheng, G., Bohm, E., Bhatele, A., Phillips, J.C., Yu, H., Kale, L.V. Scalable molecular dynamics with NAMD on the IBM Blue Gene/L System. *IBM Journal of Research and Development* **52**, 177-188 (2008).
234 Zhou, R., Berne, B. J. & Germain, R. The free energy landscape for β hairpin folding in explicit water. *Proceedings of the National Academy of Sciences* **98**, 14931-14936, doi:10.1073/pnas.201543998 (2001).
235 Walling, M. A., Novak, J. A. & Shepard, J. R. E. Quantum Dots for Live Cell and In Vivo Imaging. *International Journal of Molecular Sciences* **10**, 441-491, doi:10.3390/ijms10020441 (2009).